

**INTERNATIONAL CONFERENCE
Human-informed Translation
and Interpreting Technology**

HiT-IT 2023

P R O C E E D I N G S

Edited by Constantin Orăsan, Ruslan Mitkov,
Gloria Corpas Pastor and Johanna Monti



**International Conference on
Human-Informed Translation
and Interpreting Technology**

7-9 July 2023, Naples, Italy
<http://hit-it-conference.org>

INTERNATIONAL CONFERENCE
Human-informed Translation
and Interpreting Technology
(HiT-IT 2023)

PROCEEDINGS

7-9 July 2023, Naples, Italy
<http://hit-it-conference.org>

Series Online ISSN 2683-0078
e-book site: www.acl-bg.org

INCOMA Ltd.
Shoumen, BULGARIA

Preface

HiT-IT seeks to act as a meeting point for (and invites) researchers working in translation and interpreting technologies, practicing technology-minded translators and interpreters, companies and freelancers providing services in translation and interpreting as well as companies developing tools for translators and interpreters. In addition to the accepted papers for presentation, HiT-IT will feature invited talks by prominent experts as well as presentations and panels hosted by practitioners.

Most of the existing conferences are either focused too much on the automatic side of translation or concentrate largely on translators' and interpreters' professions. HiT-IT seeks to fill in this gap by allowing the discussion, the scientific comparison, and the mutual enrichment of professionals from both fields. HiT-IT 2023 addresses the development of translation tools and the experience translators and interpreters have with these tools as well as the development of machine translation engines, incorporating human (translators and interpreters') expertise. The conference also offers a discussion forum and publishing opportunity for professionals from the human translation and interpreting fields (e.g. translators including subtitlers, interpreters, respeakers, researchers in translation and interpreting studies) and for researchers and developers working on translation and interpreting technology and machine translation. The idea behind this conference attendees to hear the other side's position and to voice their opinions on how to make translation technologies closer to what would be accepted by large audiences, by incorporating human expertise into them.

The conference HiT-IT 2023 features seven keynote presentations:

- Jochen Hummel, Coreon,
- Constantin Orăsan, University of Surrey, UK,
- Tharindu Ranasinghe, Aston University, UK,
- Marcin Feder, Speech-to-text Unit in DG TRAD at European Parliament, Belgium,
- Manuel Herranz, Pangeanic,
- Tímea Palotai-Torzsás, Juremy.com,
- Dilyana Ilieva and Mina Ilieva, @Mitra Translations, Bulgaria.

Many thanks go to the the University of Naples L'Orientale (Italy), Lancaster University (United Kingdom), the University of Surrey (United Kingdom), the University of Malaga (Spain) and the Association of Computational Linguistics (Bulgaria).

Finally, sincere thanks to Pangeanic, Juremy.com and Mitra Translations for their generous support of HiT-IT 2023.

The Organisers

The International Conference HiT-IT 2023 is organised by:

University of Naples L'Orientale, Italy

Lancaster University, United Kingdom

University of Surrey, United Kingdom

University of Malaga, Spain

and Association of Computational Linguistics (Bulgaria)

Conference Chairs:

Gloria Corpas Pastor, University of Malaga, Spain

Ruslan Mitkov, Lancaster University, UK

Johanna Monti, University of Naples L'Orientale, Italy

Constantin Orăsan, University of Surrey, UK

Organising Committee:

Maria Pia di Buono, University of Naples l'Orientale, Italy

Dayana Abuin Rios, University of Malaga, Spain

Khadija Ait Elqih, University of Naples l'Orientale, Italy

Anastasia Bezobrazova, University of Malaga, Spain

Meriem Boulekhokh, University of Oran, Algeria

Amal El Farhmat, (University of Malaga) University of Malaga, Spain

Alfiya Khabibullina, University of Malaga, Spain

Lilit Kharatian, University of Malaga, Spain

Nikolai Nikolov, Incoma Ltd., Bulgaria

Daria Sokova, New Bulgarian University, Bulgaria

Giulia Speranza, University of Naples l'Orientale, Italy

Programme Committee:

Khetam Al Sharou, Imperial College London, United Kingdom
Eithar Alangari, Shaqra University, Saudi Arabia
Isuri Anuradha, University of Wolverhampton, United Kingdom
Silvia Bernardini, University of Bologna, Italy
Frederic Blain Tilburg, University, Netherlands
Lynne Bowker, University of Ottawa, Canada
Jon Cambra, Welocalize, Spain
Sheila Castilho, Dublin City University, Ireland
Jaleh Delfani, University of Surrey, United Kingdom
Anna Beatriz Dimas Furtado, University of Galway, Ireland
Félix do Carmo, University of Surrey United, Kingdom
Marie Escribe, Polytechnic University of Valencia / LanguageWire, Spain
Federico Gaspari, Dipartimento di Scienze Politiche, Università degli Studi di Napoli
Federico II, Italy
Amal Haddad Haddad, University of Granada, Spain
Najeh Hajlaoui, Consultant at European Commission - European Parliament, Luxembourg
Manuel Herranz, Pangeanic Spain
Maarit Koponen, University of Eastern Finland, Finland
Tomasz Korybski, University of Surrey, United Kingdom
Maria Kunilovskaya, Saarland University, Germany
Sobha Lalitha Devi, AU-KBC Research Centre, Anna University, India
Ekaterina Lapshinova-Koltunski, University of Hildesheim, Germany
Raquel Lazaro Gutierrez, Universidad de Alcalá, Spain
Judyta Mezyk, Paris-Est Créteil University, University of Silesia in Katowice, Poland
Núria Molines Galarza, Universitat Jaume I, Spain
Helena Moniz, INESC-ID, University of Lisbon, EAMT, Portugal
Joss Moorkens, Dublin City University Ireland
Elena Murgolo, Orbital14 S.r.l., Italy
David Orrego-Carmona, University of Warwick, United Kingdom
John E. Ortega, Northeastern University, United States
Tímea Palotai-Torzsás, Juremy Ltd., Hungary
Bianca Prandi, Leopold-Franzens-Universität Innsbruck, Austria
Damith Premasiri, University of Wolverhampton, United Kingdom
Ayla Rigouts Terryn, KU Leuven KULAK, Centre for Computational Linguistics, Belgium
Paola Ruffo, Ghent University, Belgium
Vilemini Sosoni, Ionian University, Greece
Maria Stasimioti, Ionian University, Greece
Elena Isabelle Tamba, Romanian Academy, Romania
Eleanor Taylor-Stilgoe, University of Surrey, United Kingdom
Irina Temnikova, Big Data for Smart Society Institute (GATE), Bulgaria, Bulgaria
Antonio Toral, University of Groningen, Netherlands
Eva Vanmassenhove, Tilburg University, Netherlands
Mihaela Vela, Saarland University (Department of Language Science and Technology), Germany
Eleni Zisi, EL-Translations, Greece

Table of Contents

<i>Gender Bias in Machine Translation: A Statistical Evaluation of Google Translate and DeepL for English, Italian and German</i> Argentina Rescigno and Johanna Monti	1
<i>The Use of Translation Technologies and Translators' Technological Competence in Saudi Arabia: Taking Stock of the past Seven Years</i> Abdelalah Alsolami, Maria Fernandez-Parra and Jun Yang	12
<i>An Exploration of Risk in the Use of MT in Healthcare Settings with Abbreviations as a Use Case</i> Eleanor Taylor-Stilgoe, Constantin Orăsan and Félix do Carmo	26
<i>Translations and Open Science: Exploring How Translation Technologies Can Support Multilingualism in Scholarly Communication</i> Susanna Fiorini, Arda Tezcan, Tom Vanallemeersch, Sara Szoc, Kristin Migdisi, Laurens Meeus and Lieve Macken	41
<i>Current Evidence of Post-edits: Differences between Post-edited Neural Machine Translation Output and Human Translation Revealed through Human Evaluation</i> Michael Farrell	52
<i>Effective Editing in Respeaking: Unveiling Uniquely Human Skills in Live Speech-to-Text Transformation</i> Tomasz Korybski and Elena Davitti	64
<i>InterpreTutor: Using Large Language Models for Interpreter Assessment</i> Cihan Ünlü	78
<i>Google Translate vs. ChatGPT: Can Non-language Professionals Trust Them for Specialized Translation?</i> Lucía Sanz-Valdivieso and Belén López-Arroyo	97
<i>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</i> Michael Farrell	108
<i>Leveraging ChatGPT and Multilingual Knowledge Graph for Automatic Post-Editing</i> Min Zhang, Xiaofeng Zhao, Zhao Yanqing, Hao Yang, Xiaosong Qiao, Junhao Zhu, Wenbing Ma, Su Chang, Yilun Liu, Yinglu Li, Minghan Wang, Song Peng, Shimin Tao and Yanfei Jiang	114
<i>The Proper Place of Men and Machines - Updated</i> Elliott Macklovitch	124
<i>Automatic Detection of Omission in Comparative Literary Translation</i> Amal Haddad Haddad	133
<i>Comparing Interface Designs to Improve RSI Platforms: In-sights from an Experimental Study</i> Muhammad Ahmed Saeed, Eloy Rodríguez González, Tomasz Korybski, Elena Davitti and Sabine Braun	147
<i>A New CAI-tool for RSI Interpreters' Training: A Pilot Study</i> Valentina Baselli	157

<i>Introducing Speech Recognition in Non-live Subtitling to Enhance the Subtitler Experience (SUBX)</i> Zeljko Radic, Sabine Braun and Elena Davitti	167
<i>Towards a Decentralize Solution for Copyrights Management in Audiovisual Translation and Media Accessibility</i> Iris Serrat-Roozen and Estella Oncins	177
<i>Developing a Customisable Subtitling Tool Based on Academic Research and User Needs</i> Zeljko Radic	188
<i>A Pedagogical Platform for Spoken Post-editing (PE): The Integration of Speech Input into COPECO</i> Jeevanthi Liyanapathirana, Pierrette Bouillon and Jonathan Mutal	195
<i>National Language Technology Platform (NLTP): The Final Stage</i> Artūrs Vasīļevskis, Jānis Ziediņš, Marko Tadić, Željka Motika, Mark Fishel, Bjarni Barkarson, Claudia Borg, Keith Aquilina and Donatienne Spiteri	203
<i>Syntactic Quality Measurement in Machine Translation with Interlinguas</i> Alessandro Maisto and Javier Oliver	209
<i>Analysis and Evaluation of ChatGPT-induced HCI Shifts in the Digitalised Translation Process</i> Pilar Sánchez-Gijón and Leire Palenzuela-Badiola	227
<i>Language Complexity in Human and Machine Translation: A Preliminary Study</i> Gabor Recski and Fanni Kadar	268
<i>Human Evaluation for Translation Quality of ChatGPT: A Preliminary Study</i> Zhao Yanqing, Min Zhang, Xiaoyu Chen, Yadong Deng, Aiju Geng, Limin Liu, Xiaoqin Liu, Wei Li, Yanfei Jiang, Hao Yang, Yu Han, shimin tao, Xiaochun Li, Ma Miaomiao, Zhaodi Zhang and Xie Ning	282
<i>Comparing MTQE Scores with Fuzzy Match Percentages from CAT Tools</i> Elena Murgolo	288
<i>Are Post-editeese Features Really Universal?</i> Lise Volkart and Pierrette Bouillon	294
<i>Behind the Scenes: Freelance Translators Use of Machine Translation in the Medical Field</i> Magali Vidrequin	305
<i>Workbench for Post-editing of Translations from English to Dravidian Languages</i> Sobha Lalitha Devi, Pattabhi R K Rao and Vijay Sundar Ram	315
<i>Processing English Verb Phrase Ellipsis for Conversational English-Hindi Machine Translation</i> Aniruddha Deshpande and Dipti Sharma	325
<i>Lost in Innu-Aimun Translation - Re-defining NMT for Indigenous Interpreters and Translators Needs</i> Antoine Cadotte, Anne-Christina Thernish and Fatiha Sadat	342

Conference Program

Friday, July 7, 2023

09:15–09:30 *Welcome and opening of the conference*

09:30–11:00 **Keynote/Sponsor’s presentation**

09:30–10:30 *Embracing LangOps: Mandatory in the Generative AI Era*
Jochen Hummel, CEO Coreon

10:30–11:00 *sTMS Cloud – a Boutique Translation Project Management System by Mitra Translations*
Dilyana Ilieva, CEO, Mitra Translations and Mina Ilieva, Managing Director, Mitra Translations

11:30–13:00 **Long presentations session 1: Machine translation**

11:30–12:00 *Gender Bias in Machine Translation: A Statistical Evaluation of Google Translate and DeepL for English, Italian and German*
Argentina Rescigno and Johanna Monti

12:00–12:30 *The Use of Translation Technologies and Translators’ Technological Competence in Saudi Arabia: Taking Stock of the past Seven Years*
Abdelalah Alsolami, Maria Fernandez-Parra and Jun Yang

12:30–13:00 *An Exploration of Risk in the Use of MT in Healthcare Settings with Abbreviations as a Use Case*
Eleanor Taylor-Stilgoe, Constantin Orăsan and Félix do Carmo

Friday, July 7, 2023 (continued)

15:00–15:45 Sponsor’s presentation

15:00–15:45 *The Importance is Data Anonymization to Build Ethical AI*
Manuel Herranz, CEO, Pangeanic

15:45–16:25 Short presentations session 1: Machine translation

15:45–16:05 *Translations and Open Science: Exploring How Translation Technologies Can Support Multilingualism in Scholarly Communication*
Susanna Fiorini, Arda Tezcan, Tom Vanallemeersch, Sara Szoc, Kristin Migdisi, Laurens Meeus and Lieve Macken

16:05–16:25 *Current Evidence of Post-edits: Differences between Post-edited Neural Machine Translation Output and Human Translation Revealed through Human Evaluation*
Michael Farrell

16:50–17:50 Long presentations session 2: Interpreting and respeaking

16:50–17:20 *Effective Editing in Respeaking: Unveiling Uniquely Human Skills in Live Speech-to-Text Transformation*
Tomasz Korybski and Elena Davitti

17:20–17:50 *InterpreTutor: Using Large Language Models for Interpreter Assessment*
Cihan Ünlü

Saturday, July 8, 2023

09:30–10:30 Keynote presentation

09:30–10:30 *GenAI: the next step in the evolution of translation?*
Constantin Orasan, University of Surrey, UK

10:30–11:00 Long presentations session 2: ChatGPT

10:30–11:00 *Google Translate vs. ChatGPT: Can Non-language Professionals Trust Them for Specialized Translation?*
Lucía Sanz-Valdivieso and Belén López-Arroyo

11:30–12:00 Sponsor’s presentation

11:30–12:00 *Smart concordance search on EU terminology integrated into Trados Studio – the Juremy Web Plugin*
Tímea Palotai-Torzsás, Co-Founder of Juremy.com

12:00–13:00 Short presentations session 2: ChatGPT and translation

12:00–12:20 *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*
Michael Farrell

12:20–12:40 *Leveraging ChatGPT and Multilingual Knowledge Graph for Automatic Post-Editing*
Min Zhang, Xiaofeng Zhao, Zhao Yanqing, Hao Yang, Xiaosong Qiao, Junhao Zhu, Wenbing Ma, Su Chang, Yilun Liu, Yinglu Li, Minghan Wang, Song Peng, shimin tao and Yanfei Jiang

12:40–13:00 *The Proper Place of Men and Machines - Updated*
Elliott Macklovitch

Saturday, July 8, 2023 (continued)

15:00–16:00 Keynote presentation

15:00–16:00 *Towards Responsible Machine Translation with Quality Estimation*
Tharindu Ranasinghe, Aston University, UK

16:00–16:30 Long presentations session 3: Machine translation

16:00–16:30 *Automatic Detection of Omission in Comparative Literary Translation*
Amal Haddad Haddad

17:00–18:20 Short presentations session 3: Interpreting

17:00–17:20 *Comparing Interface Designs to Improve RSI Platforms: In-sights from an Experimental Study*
Muhammad Ahmed Saeed, Eloy Rodríguez González, Tomasz Korybski, Elena Davitti and Sabine Braun

17:20–17:40 *A New CAI-tool for RSI Interpreters' Training: A Pilot Study*
Valentina Baselli

17:40–18:00 *Introducing Speech Recognition in Non-live Subtitling to Enhance the Subtitler Experience (SUBX)*
Zeljko Radic, Sabine Braun and Elena Davitti

18:00–18:20 *Towards a Decentralize Solution for Copyrights Management in Audiovisual Translation and Media Accessibility*
Iris Serrat-Roozen and Estella Oncins

Sunday, July 9, 2023

09:30–10:30 Keynote presentation

09:30–10:30 *Live speech-to-text and machine translation tool for 24 languages – increasing accessibility of parliamentary debates*
Marcin Feder, Head of the Speech-to-text Unit in DG TRAD

10:30–11:00 Demo booster session + demos

10:30–10:40 *Developing a Customisable Subtitling Tool Based on Academic Research and User Needs*
Zeljko Radic

10:40–10:50 *A Pedagogical Platform for Spoken Post-editing (PE): The Integration of Speech Input into COPECO*
Jeevanthi Liyanapathirana, Pierrette Bouillon and Jonathan Mutal

10:50–11:00 *National Language Technology Platform (NLTP): The Final Stage*
Artūrs Vasīļevskis, Jānis Ziedīņš, Marko Tadić, Željka Motika, Mark Fishel, Bjarni Barkarson, Claudia Borg, Keith Aquilina and Donatienne Spiteri

11:30–13:00 Long presentations session 4: Machine translation

11:30–12:00 *Syntactic Quality Measurement in Machine Translation with Interlinguas*
Alessandro Maisto and Javier Oliver

12:00–12:30 *Analysis and Evaluation of ChatGPT-induced HCI Shifts in the Digitalised Translation Process*
Pilar Sánchez-Gijón and Leire Palenzuela-Badiola

12:30–13:00 *Language Complexity in Human and Machine Translation: A Preliminary Study*
Gabor Recski and Fanni Kadar

Sunday, July 9, 2023 (continued)

15:00–16:20 User papers and short presentations session 4: Machine translation

15:00–15:20 *Human Evaluation for Translation Quality of ChatGPT: A Preliminary Study*
Zhao Yanqing, Min Zhang, Xiaoyu Chen, Yadong Deng, Aiju Geng, Limin Liu,
Xiaoqin Liu, Wei Li, Yanfei Jiang, Hao Yang, Yu Han, shimin tao, Xiaochun Li, Ma
Miaomiao, Zhaodi Zhang and Xie Ning

15:20–15:40 *Comparing MTQE Scores with Fuzzy Match Percentages from CAT Tools*
Elena Murgolo

15:40–16:00 *Are Post-edited Features Really Universal?*
Lise Volkart and Pierrette Bouillon

16:00–16:20 *Behind the Scenes: Freelance Translators Use of Machine Translation in the Medi-
cal Field*
Magali Vidrequin

16:50–17:50 Short presentations session 5: Low resource languages

16:50–17:10 *Workbench for Post-editing of Translations from English to Dravidian Languages*
Sobha Lalitha Devi, Pattabhi R K Rao and Vijay Sundar Ram

17:10–17:30 *Processing English Verb Phrase Ellipsis for Conversational English-Hindi Machine
Translation*
Aniruddha Deshpande and Dipti Sharma

17:30–17:50 *Lost in Innu-Aimun Translation - Re-defining NMT for Indigenous Interpreters and
Translators Needs*
Antoine Cadotte, Anne-Christina Thernish and Fatiha Sadat

Gender Bias in Machine Translation: a statistical evaluation of Google Translate and DeepL for English, Italian and German

Argentina Anna Rescigno¹[0009-0000-3653-8492] and Johanna Monti¹[0000-0002-4563-5988]

UNIOR NLP Research Group University of Naples "L'Orientale"
a.rescigno1@studenti.unior.it
jmonti@unior.it

Abstract. Despite the significant advancements made in the field of Machine Translation (MT) technology, there are still some challenges that need to be addressed. One such challenge is represented by the issue of gender bias in machine translation systems. The main objective of this study is to examine and investigate the presence of gender bias in MT systems and identify any potential issues related to the use of sexist language. The research evaluates the performance of Google Translate and DeepL in terms of natural gender translation, particularly the frequency of male and female forms used in translating sentences that refer to professions without any other gender-specific words. The evaluation is carried out using the MT-GenEval corpus [2] contextual subset, for English-Italian and English-German language pairs. The paper presents the statistical findings obtained from the evaluation.

Keywords: Machine Translation · Gender Bias · Natural Language Processing.

1 Introduction

Bias in automated systems refers to the tendency of these systems to repeatedly make the same assumptions. In the scientific community, bias in artificial intelligence (AI) has become a significant concern due to the increasing use of AI applications such as machine learning architectures. These systems learn by maximizing prediction accuracy, which means that they optimise themselves based on patterns that appear more frequently in their training data. However, if a certain phenomenon is overrepresented in the training data, the program will optimize for it, as this increases its accuracy [14]. The algorithms are evaluated on sub-samples of original training sets making them more likely to exhibit the same biases observed during training. This recurring biased behaviour is responsible for the lack of diversity in machine translation outputs on multiple levels, which can be attributed to the observed algorithmic bias [12].

In 2005, Koehn [3], identified in a preliminary analysis of the Europarl corpus a male-to-female speakers ratio of 2:1. The corpus is made of about 30 million

words for each of the 11 official languages of the European Union (Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese, and Swedish) and it is one of the most used corpora to train MT systems. Consequently, masculine noun endings or pronouns are more frequently used in some languages where gender agreement with the speaker is required, leading to a bias in translation. As a matter of fact, this translation bias reinforces the existing gender disparity and further exacerbates it.

Nonetheless, not until recent times, limited attention has been given to gender bias in machine translation, specifically regarding the natural gender of languages [10]. One way to identify examples of biases in machine translation is to examine an ambiguous sentence in the source language that is unambiguous in the target language. Ambiguity can occur when a linguistic feature is explicit in one language and not in another one or vice-versa. For instance, in English, natural gender is implicitly understood, while in languages such as Italian and German, it must be grammatically expressed. For example, in a simple sentence such as “*I am happy*”, the subject “*I*” has an ambiguous gender in English. However, in Italian, the gender must be grammatically expressed as “*Io sono contento*” for a male subject and “*Io sono contenta*” for a female subject. Typical users of machine translation systems may not be aware of this bias in the machine translation outputs, especially if they are not proficient in the source language, and there is currently no tool to notify them about it. However, those who are familiar with the target language may recognise when a specific gender is being used inappropriately or offensively. To mitigate gender bias in machine translation, gender characteristics can be integrated into the training data for neural machine translation systems. However, these types of approaches are still limited and further investigation is needed.

The majority of modern machine translation systems translate at the sentence level, meaning that gender-related issues are usually addressed through statistical methods using training data. As a result, translation errors relating to gender are more common in translations between languages, such as English, which retains features relating to natural gender only with some nouns and pronouns and neuter pronouns for sexless objects, and morphologically rich languages like Italian, French, Spanish, and German. These latter languages require additional information in order to accurately translate gender. When this information is missing, the systems tend to generate the most frequent variants, often resulting in biased translations. A study conducted by Prates et al. [6] found that Google Translate defaults to producing more male-gendered outputs than expected, even when taking into account demographic data. This suggests the existence of a phenomenon known as machine bias [6] [13].

2 Related Works

Several studies have been carried out in the field of Machine Translation to address gender bias in translation outputs. These studies involve training trans-

lation systems, developing specific datasets to mitigate gender bias, and creating external tools to integrate into the MT systems themselves.

Monti [4] carried out a study to explore the issues related to the translation of gender evaluating four MT systems - i.e., Google Translate, Microsoft BING Translator, Lucy KWIK Translator and SYSTRANet. While the first two systems adopt a neural approach, Lucy is a rule-based system and SYSTRAN combines both statistical and linguistic approaches. Monti’s study explored various types of issues in gender translation, such as subject-nominate predicate agreement (e.g., *Mary is a diligent employee*), subject-object agreement (e.g., *Jane seems very nervous*), name-apposition agreement (e.g., *Mary, our doctor, went away a few minutes ago*), name-past participle agreement (e.g., *Mary is happily married*), name-anaphoric/cataphoric reference (e.g., *The student studied really hard for her test*). The study revealed that all approaches had translation errors regarding gender.

Despite advancements in the field of Natural Language Processing (NLP), the issue of gender translation continues to be a significant concern, as highlighted by Monti [5]. This issue was further exemplified by a recent study conducted by Rescigno et al. [8] which evaluated the frequency of feminine, masculine and neutral forms in the translation of nouns referring to professions when related to certain adjectives (i.e., *beautiful, efficient, intelligent, sad, famous*) using three state-of-the-art MT systems. The evaluation focused on English–Italian, English–Spanish and English–French language pairs.

Interlinguistic differences between languages require that implicit information from the source language be made explicit in the target language. Human translators can consider the context (both linguistic and extra-linguistic) to infer any necessary information and translate accordingly, while MT systems may encounter challenges in accomplishing this task. Vanmassenhove and Monti [11] developed gENDER-IT, a collection of annotated sentences that focuses on gender-related phenomena, to address the lack of high-quality datasets tailored to investigate specific interlinguistic phenomena, including gender and/or number. The dataset was manually adapted from the MuST-SHE Corpus [1] but, unlike this one, gENDER-IT tags also include - other than masculine/feminine referents - a neutral referent, whereas the sentence itself does not provide any explicit hint to detect the gender. The dataset indeed focuses on ambiguous English sentences providing the correct translations in Italian. For example, the English sentence *"Do you remember that patient you sent home?" the other nurse asked matter-of-factly* can, in fact, have four different translation alternatives, all correct:

- 1) *"Si ricorda **quel** paziente (M) che ha mandato a casa?" mi ha chiesto l'altro infermiere (M);*
- 2) *"Si ricorda **quel** paziente (M) che ha mandato a casa? " mi ha chiesto l'altra infermiera (F);*
- 3) *"Si ricorda **quella** paziente (F) che ha mandato a casa? " mi ha chiesto l'altra infermiera (F);*
- 4) *"Si ricorda **quella** paziente (F) che ha mandato a casa? " mi ha chiesto l'altro infermiere (M).*

Finally, more recent works focusing on the creation of specific datasets, tailored to evaluate the accuracy of MT outputs according to gender, include the one from Currey et al. [2], MT-GenEval. This set is based on real-world data from Wikipedia and covers the translations from English towards 8 target languages (Arabic, French, Hindi, Italian, Portuguese, Russian, Spanish, and German). Moreover, this set integrates existing benchmarks, such as WinoMT,¹ MuST-SHE corpus [1], and GeBioCorpus,² providing realistic and gender-balanced counterfactual data for all the languages considered. Further evaluations which include more linguistic combinations and even more centred and balanced regarding gender are presented by Rarrick et al. [7] with the GATE corpus (*Gender-Ambiguous Translation Examples*). GATE is a linguistically diverse corpus consisting of gender-ambiguous source sentences in English along with multiple alternative translations in the target languages considered, i.e., Spanish, French and Italian.³ The corpus is made up of thousands of segments which consist of a single English sentence paired with one or two alternative translations in Spanish, French or Italian; each sentence contains at least an *Arbitrarily Gender-Marked Entity* (AGME), which results in an unmarked name referring to an animate entity in the source language (English) but marked for gender in the target languages.

3 Materials and Methods

This study employs a statistical approach to analyse gender bias in machine translation. Its main objective is to detect instances of gender stereotypes in widely used MT systems, Google Translate and DeepL, for English–German and English–Italian language pairs. Furthermore, the present study aims to examine whether the utilization of an extended context, as proposed by Tiedemann [9], specifically through the incorporation of extended translation units encompassing the sentence under investigation and the preceding sentence, yields any positive or negative effects in the disambiguation and translation of gender, particularly when the gender of the referent or speaker is ambiguous. For example, in the sentence “*Godoy is an accountant by profession and is currently studying to receive a degree in social work*”, the gender of the entity described with the English substantive “*accountant*” can only be accurately identified by considering the context, namely the previous sentence, from which it is evident that it is referring to a female entity due to the presence of the pronoun “*her*”:

<context> *At the age of 17, Godoy was assaulted by a police officer while protesting the Augusto Pinochet dictatorship and was put in prison for **her** social activism.* <sentence> *Godoy is an **accountant** by profession and is currently studying to receive a degree in social work.*

¹ https://github.com/gabrielStanovsky/mt_gender

² <https://github.com/PLXIV/Gebiotoolkit>

³ All data, including the corpus and the specific evaluation tool, are publicly available at Github repository <https://github.com/MicrosoftTranslator/GATE>

3.1 Description of the Dataset

The study utilized the MT-GenEval dataset (*Machine Translation Gender Evaluation* benchmark), which is a reference dataset created specifically to assess the accuracy of machine translation outputs regarding gender. Unlike other datasets involved in bias identification research, the MT-GenEval dataset does not use synthetic data, but it is based on authentic data from Wikipedia. It contains realistic counterfactual data that is gender-balanced across all included languages. The dataset covers 8 language pairs, consisting of a total of 2,400 sentences each.⁴

For this evaluation, the contextual dataset was considered, which includes sentences containing at least one profession noun without any other phrasal element that could indicate the gender of the referent. These selected sentences are accompanied by one or two antecedent contextual sentences that aim to clarify the gender of the selected referent.

Consequently, the dataset consists of sentences that are ambiguous in terms of the gender of the referent but have a previous contextual sentence that clarifies the gender information (i.e., contains gender information). The sentence set was manually edited, removing any unclear or excessively ambiguous examples, for example

<context> *Cook and Dickerman made this their home and Eleanor had her own room, although she rarely spent the night.* <sentence> *Cook, an expert woodworker, made all furniture.*

Also, some other sentences lacking in the contextual information have been removed as well, such as:

<context> ... <sentence> *A surrogate's life may be very similar to that of the author.*

<context> ... <sentence> *Cook was pressed into service as a player as a result of injuries.*

The final dataset used for the evaluation includes 448 sentences in Italian and 557 sentences in German.

3.2 Experimental Setup

The methodology used for this research can be divided into two phases. In the first phase, translations of single sentences, i.e. without context, obtained from Google Translate and DeepL for both the Italian and German language sets were evaluated manually. The translations were tagged based on the gender of the referent, which could be either masculine (M), feminine (F), or "non-available" (N/A) when the gender could not be inferred. Subsequently, the outputs from the machine translation systems were compared to the benchmark translations, which consisted of the original sentences from the MT-GenEval corpus. The MT system outputs were manually tagged as either "positive correspondence" (Y), "negative correspondence" (N), or "ambiguous" (A) when the gender of

⁴ <https://github.com/amazon-science/machine-translation-gender-eval>

the referent was not clear. In this way, a statistic on the percentage of male and female outputs that matched the gender of the benchmark sentences has been obtained.

The second part of the experiment focused on examining the sentences again by reintroducing them into the machine translation (MT) systems along with the sentences containing the contextual information to ascertain any changes in the outputs. Similarly, also these latter results have been gender tagged and, after being compared to the benchmark sentences, their correspondence regarding the gender of the referent has been annotated as well.

All the machine translations and related manual evaluations were conducted between February and March of 2023 and are accessible in the publicly available GitHub repository.⁵

4 Results and Analysis

The gender-annotated sets for Italian and German languages present the following results (Table 1), which will be hereinafter indicated as *Benchmark* (BM).

Table 1. “Clean” dataset statistics (Benchmark).

ITALIAN			GERMAN		
F	M	Total	F	M	Total
199	249	448	245	312	557
44,4%	55,6%	100%	44%	56%	100%

After the manual input of the sentences into the MT systems under consideration, the obtained results were assessed based on their level of correspondence, with the gender of the referent as presented in the benchmark translation.

A preliminary analysis reveals that the majority of translation outputs have a masculine referent in both the MT systems. Specifically, Google Translate produces about 90% male outputs for both languages; DeepL generates 85% male outputs for Italian and 88% for German.

It is worth noting that there were relatively fewer benchmark sentences with masculine gender compared to the MT outputs (Google Translate produces 406 male-gendered sentences for Italian and 506 male-gendered sentences for German in comparison with the benchmark which includes 249 male-gendered sentences for Italian and 312 male-gendered sentences for German; similarly, DeepL produces 385 male-gendered sentences for Italian and 492 male-gendered sentences for German) (Table 2), making the higher production of masculine translation outputs particularly noticeable. In contrast, the number of feminine-gendered outputs is considerably low for both systems: on average, Google Translate produces 5% female-gendered output results, while DeepL has an average of 8%

⁵ https://github.com/argentina-res/genderbias_dissertation

female-gendered outputs. From Table 2, it emerges very clearly that both MT systems are biased, as they show a higher percentage of masculine gender in comparison with the benchmark translations in Italian and German.

Besides distinguishing translation results between male (M) and female (F) genders - and, neutral (Neut) for what concerns German language - a further differentiation was made for some translated sentences where there was no other useful element for defining the gender of the referent; this category includes sentences tagged as N/A.

Table 2. Results in number (#) and % of masculine (M), feminine (F) and non-available gender (N/A) outputs (also, neutral (N) translation outputs for German) obtained from Google Translate (GT) and DeepL (DL) MT systems for the Italian (IT) and German (DE) languages in relation to the Benchmark (BM) translation statistics.

		BM	%BM	#GT	%GT	#DL	%DL
IT	M	249	55,6	406	90,6	385	85,9
	F	199	44,4	19	4,3	36	8
	N/A	-	-	23	5,1	27	6
Total		448	100	448	100	448	100
DE	M	312	56	506	90,8	492	88,3
	F	245	44	34	6,1	55	9,9
	N/A	-	-	14	2,5	7	1,3
	Neut	-	-	3	0,5	3	0,5
Total		557	100	557	100	557	100

The evaluation process entailed comparing the outputs of the machine translation (MT) systems with the benchmark translations. A systematic analysis was performed for each MT output to ascertain and classify the correspondence between the gender of the sentence’s referent and the gender of the referents in the benchmark sentences, thereby assigning either a positive (Y) or negative (N) tag. In cases where this kind of correspondence cannot be identified, the ambiguous (A) tag has been employed.

Table 3. Results in number (#) and % of cases of positive (Y), negative (N) and ambiguous (A) correspondence between benchmark translations and sentences without context translated by the MT systems Google Translate (GT) and DeepL (DL) for the Italian (IT) and German (DE) languages.

		#GT	%GT	#DL	%DL
IT	Y	254	56,7	275	61,4
	N	173	38,6	146	32,6
	A	21	4,7	27	6
DE	Y	333	59,8	363	65,2
	N	214	38,4	187	33,6
	A	10	1,8	7	1,2

Table 3 presents the results of the evaluation of the correspondence between the MT outputs and the benchmark translations, indicating the number of positive, negative and ambiguous cases accompanied by their respective percentages. The findings demonstrate a relatively equal distribution between the two systems and the considered languages, regarding their negative and positive correspondence. Google Translate, in fact, shows always more than 50% positive correspondence (Y) with the benchmark sentences for both language pairs (56.7% regarding Italian, 59.8% for German), whereas DeepL presents a percentage above 60% (61.4% for Italian and 65% for German).

In regards to cases where the translated output does not correspond in terms of gender with the benchmark translation, the percentages of such cases are below 40% for both systems: Google Translate achieves approximately 38% negative results for both language pairs, while DeepL results range between 32% and 33%.

At this stage, the experiment includes adding antecedent contextual sentences to all the single sentences under examination in order to detect any positive or negative changes in the MT output for what concerns gender disambiguation. Subsequently, those results have been compared with the respective MT outputs without context.

The findings for Google Translate and DeepL for both the English-Italian and English-German language pairs (Fig. 1) show that the number of male outputs obtained after translating the sentences with the contextual sentences differs from the previous results concerning the translation of the sentences without context (Table 2): in particular, the number of feminine and N/A outputs increases. What differs between the two systems is the different percentages of masculine, feminine, and non-available gender outputs, resulting in DeepL outperforming Google Translate, as it produces less masculine defaults.

Subsequently, the aforementioned procedure was repeated, where the translated sentences, along with their corresponding context, were compared to the benchmark sentences.

As Fig. 1 shows, it is evident that the benchmark sentences exhibit higher percentages of positive correspondence with respect to the gender of the referents in the benchmark sentences. Once again, Google Translate demonstrates a marginal improvement compared to DeepL, with the disparity between the "before" and "after" stages of incorporating contextual information becoming more pronounced.

However, the findings also reveal some instances in which the inclusion of contextual information not only fails to enhance the performance of the machine translation (MT) systems but instead has the opposite effect. This occurs due to an erroneous disambiguation of the gender of the referent, a situation that had not previously arisen despite context was not provided. However, such occurrences are relatively infrequent, as evidenced by the data presented in Fig. 2. In the case of Google Translate, there were only a total of 6 instances where the correct gender disambiguation deteriorated for the English-Italian language pair and 16 instances for English-German translations. Conversely, DeepL presents a

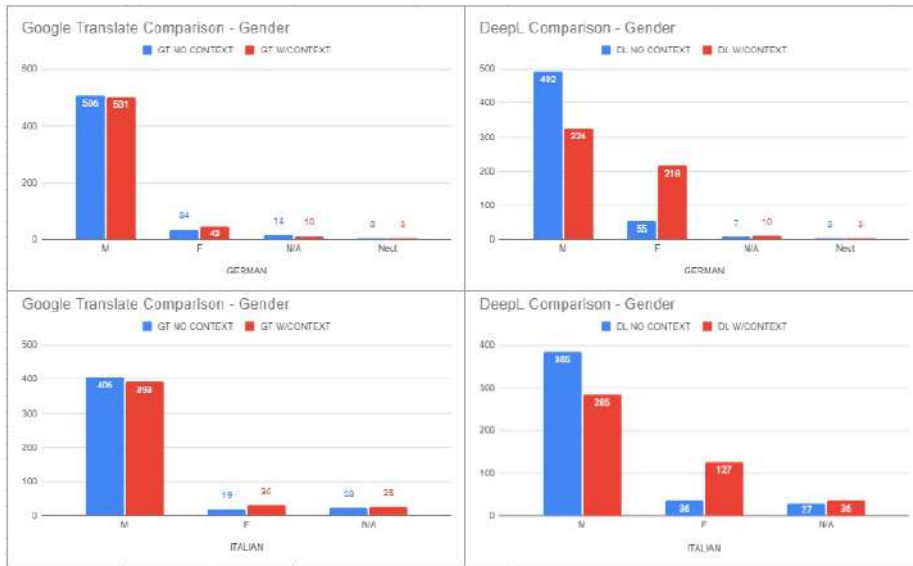


Fig. 1. Graphs showing a comparison of the number of gendered outputs resulting for both Google Translate and DeepL before (*no context*) and after (*w/context*) adding the contextual information for English-German translations (*upper graphs*) and English-Italian translations (*lower graphs*).

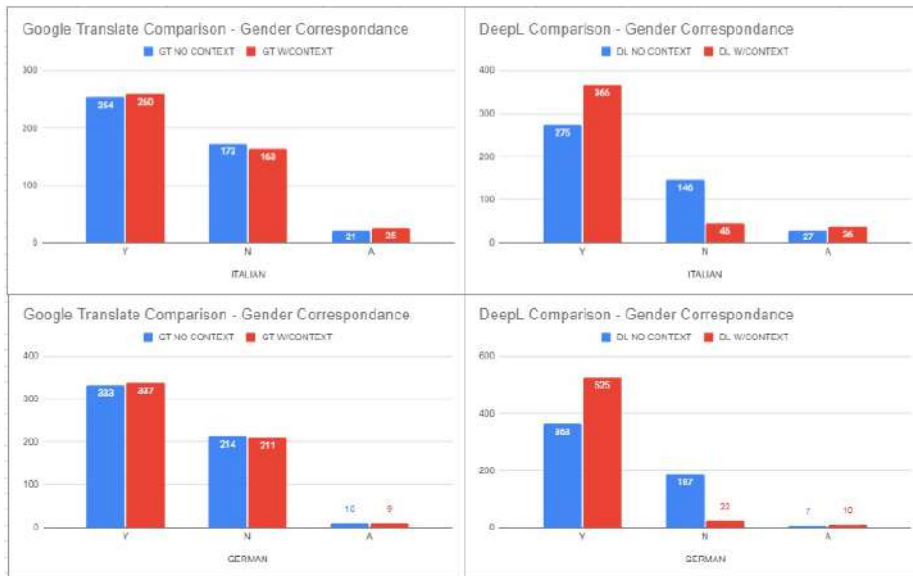


Fig. 2. Graphs showing a comparison of the number of positive (Y), negative (N), and ambiguous (A) concordance results for both Google Translate and DeepL before (*no context*) and after (*w/context*) adding the contextual information for English-Italian translations (*upper graphs*) and English-German translations (*lower graphs*).

lower number of cases for both language pairs - respectively, 3 for Italian and 4 for German.

5 Conclusions and Future Work

The present study involved comparing two machine translation (MT) systems for the English-Italian and English-German language pairs by examining the frequency of male and female gender usage.

The findings of this study indicate, that the MT systems still have a tendency to default to the masculine gender, with some slight differences between Google Translate and DeepL. Moreover, the inclusion of contextual sentences demonstrates overall improvements in the results concerning the positive correspondence of gender agreement with the benchmark translations. Finally, comparing the two system outputs, it is clear that DeepL takes more advantage of the context sentences than Google Translate as it outperforms the latter in increasing its percentage of positive correspondences with the benchmark translations both for the Italian and German languages. However, it is also interesting to highlight that the integration of contextual information also produces some cases of incorrect gender disambiguation, whereas the gender was correctly identified in the first phase of the experiment. Also, in this case, DeepL presents a lower number of degraded occurrences.

Nevertheless, this investigation presents some limitations, including its exclusive focus on binary gender categories - masculine and feminine. The examination of binary gender within current society might be considered restrictive considering the prevailing trend towards increasing inclusivity, encompassing various aspects such as linguistic inclusivity. As a matter of fact, this and other related studies could lay the groundwork for future research endeavours that encompass a broader range of gender identities.

As a potential future direction for this research methodology, there is also the possibility of conducting more comprehensive evaluations that encompass a wider range of language combinations and more targeted and balanced datasets, especially regarding gender - such as the very recent GATE (Gender-Ambiguous Translation Examples) [7].

References

1. Bentivogli, L., Savoldi, B., Negri, M., Di Gangi, M.A., Cattoni, R., Turchi, M.: Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE corpus. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 6923–6933. Association for Computational Linguistics, Online (Jul 2020). <https://doi.org/10.18653/v1/2020.acl-main.619>
2. Currey, A., Nadejde, M., Pappagari, R.R., Mayer, M., Lauly, S., Niu, X., Hsu, B., Dinu, G.: MT-GenEval: A Counterfactual and Contextual Dataset for Evaluating Gender Accuracy in Machine Translation. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. pp. 4287–4299. Association

- for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022), <https://aclanthology.org/2022.emnlp-main.288>
3. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: *Proceedings of Machine Translation Summit X: Papers*. pp. 79–86. Phuket, Thailand (Sep 13-15 2005), <https://aclanthology.org/2005.mtsummit-papers.11>
 4. Monti, J.: Questioni di Genere in Traduzione Automatica. *Al femminile. Scritti linguistici in onore di Cristina Vallini* **139**, 411–431 (2017)
 5. Monti, J.: Gender Issues in Machine Translation: An Unsolved Problem? In: *The Routledge Handbook of Translation, Feminism and Gender*, pp. 457–468. Routledge (2020)
 6. Prates, M.O., Avelar, P.H., Lamb, L.C.: Assessing Gender Bias in Machine Translation: a Case Study With Google Translate. *Neural Computing and Applications* **32**, 6363–6381 (2020)
 7. Rarrick, S., Naik, R., Mathur, V., Poudel, S., Chowdhary, V.: Gate: A Challenge Set for Gender-Ambiguous Translation Examples. arXiv preprint arXiv:2303.03975 (2023)
 8. Rescigno, A.A., Vanmassenhove, E., Monti, J., Way, A.: A Case Study of Natural Gender Phenomena in Translation: a Comparison of Google Translate, Bing Microsoft Translator and DeepL for English to Italian, French and Spanish. *Computational Linguistics CLiC-it 2020* p. 359 (2020)
 9. Tiedemann, J., Scherrer, Y.: Neural Machine Translation with Extended Context. arXiv preprint arXiv:1708.05943 (2017)
 10. Vanmassenhove, E.: On the Integration of Linguistic Features into Statistical and Neural Machine Translation. arXiv preprint arXiv:2003.14324 (2020)
 11. Vanmassenhove, E., Monti, J.: gENDER-IT: An Annotated English-Italian Parallel Challenge Set for Cross-Linguistic Natural Gender Phenomena. In: *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. pp. 1–7. Association for Computational Linguistics, Online (Aug 2021). <https://doi.org/10.18653/v1/2021.gebnlp-1.1>, <https://aclanthology.org/2021.gebnlp-1.1>
 12. Vanmassenhove, E., Shterionov, D., Gwilliam, M.: Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. arXiv preprint arXiv:2102.00287 (2021)
 13. Vanmassenhove, E., Shterionov, D., Way, A.: Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In: *Proceedings of Machine Translation Summit XVII: Research Track*. pp. 222–232. European Association for Machine Translation, Dublin, Ireland (Aug 2019), <https://aclanthology.org/W19-6622>
 14. Zou, J., Schiebinger, L.: Ai Can be Sexist and Racist—it’s Time to Make it Fair (2018)

The Use of Translation Technologies and Translators' Technological Competence in Saudi Arabia: Taking Stock of the Past Seven Years

Abdelalah Alsolami^{1,2}[0009-0001-0366-326X], Maria Fernandez Parra¹[0000-0001-7493-0508] and Jun Yang¹[0000-0002-0004-0956]

¹ Swansea University, Singleton Park SA2 8PP, Wales, UK
943760@swansea.ac.uk

² University of Jeddah, Jeddah, Saudi Arabia

Abstract. The Saudi Vision 2030 has dramatically revolutionized the translation market in Saudi Arabia and focused on two main principles, i.e., professionalization and employability. In globalized translation markets, using translation technologies is no longer a luxury but a necessity for professional translators to respond to the increasing demand for translation services. Therefore, one research area worth looking at and investigating is the professional aspect of the Saudi translation market in terms of the technological competence of translators. This paper is a first step towards working out the technological competence of professional translators in Saudi Arabia in future by taking stock of the last seven years. In this paper, professional translators were surveyed to learn about their views of translation technology training provided by Saudi academic programmes and to self-assess their competency levels in different areas related to translation technologies. The results indicate that although the majority of respondents confirmed their high level of technological competence, some professional translators still need additional tools training to ensure that they are prepared for the demands of the modern translation market. The study has implications for university training programmes to make adjustments in their curricula to include sufficient training on translation technologies. The study also suggests that there might be a need to reconsider the concept of technological competence in light of the recent developments of translation technologies and tools as well as the increasing role of machines in the translation process.

Keywords: Translation Technologies, Technological Competence, Translator Training, Competence assessment.

1 Introduction

Translation technologies have become increasingly popular among professional translators as part of the translation process in today's globalized markets. Bowker and Pastor (2015) point out that translators could increase their productivity and income and preserve the efficiency and consistency of their translations thanks to the technologies used in their work. Moreover, it has been possible for translators to use technology to

work on collaborative translation projects and/or work remotely from across the world. In this sense, O'Hagan (2019) illustrates that computer-assisted translation (CAT) tools and machine translation (MT) have gradually transformed the human role in the translation process, moving to a machine-centered activity assisted by the interventions of humans in an interactive way, as Olohan (2011:343) puts it "a dance of agency". As technological advances, (e.g., the launch of neural machine translation NMT, the emergence of new CAT tool functionalities and products) are on the rise, keeping up with the evolving landscape of translation technology requires a sufficient level of technological competence on the part of professional translators to remain relevant and competitive in the translation industry. Previous studies have mentioned that empirical research on translation competence (TC) is still scarce and fragmented in terms of investigating the extent to which translation technologies are adopted among professional translators and their technological competence level (e.g., Olohan, 2011) especially in growing translation markets, such as the case of Saudi Arabia. As part of a broader research project, the present work reports the results of an empirical survey distributed to professional translators working in the Saudi translation market to investigate their use of translation technologies and to assess the status of their technological competence.

1.1 Context and Rationale of the Study

The current study is driven by two main motives. First, translation technologies have been notably improved during the past seven years (April 2016 – April 2023), e.g., MT improvements and MT integration into the functionality of CAT tools. However, empirical research seems to be slower than the rapid technological improvements in the translation industry; thereby, this necessitates making periodic studies to reflect the reality of the translation market in terms of investigating the use of translation technologies and tools and monitoring the status of translators' technological competence. Second, the Saudi translation market has received generous governmental support and witnessed some regulatory measures since the launch of the Saudi Vision 2030¹ in April 2016, which has the potential to entail some demographic, administrative, and/or professional changes in the market. To illustrate that, the Ministry of Culture (MoC) was established in 2018 to play a significant role in the development and growth of the local translation market in Saudi Arabia. Two years later, MoC founded the Literature, Publishing, and Translation Commission² to be a responsible authority to regulate and manage the translation sector in Saudi Arabia. This commission has announced several translation initiatives which are expected to result in a growing demand for high-quality translations. Moreover, the Ministry of Human Resources and Social Development has been working towards achieving a strategic goal of the vision: to decrease the unemployment rate among Saudi nationals to 7% by the end of the decade. In this direction, the Ministry has implemented an employment measure in the translation market since

¹ <https://www.vision2030.gov.sa/>

² The Literature, Publishing, and Translation Commission was established in February 2020
<https://lpt.moc.gov.sa/en/lpt-homepage>

May 2022, that is, to restrict translator jobs to Saudi citizens as part of a process called ‘Saudization’. This process aims to minimize the unemployment rate among Saudis and ensure that translation job opportunities are primarily available to them.

Undoubtedly, such changes will, directly and indirectly, impact the Saudi translation market in the next few years. Locally, it is more likely that these enhancements increase the demand for translation services and the level of competitiveness among Saudi translators to keep up with the needs of the growing translation market. Moreover, the enhancements are expected to increase the pressure on the national workforce, i.e., professional translators, to show high-level performance and skills to satisfy their translation employers and customers. Globally, the country's openness to foreign investments and its globalization and internationalization efforts could increase the demand for translation and localization services and help make the translation market in Saudi Arabia attractive for more translation works from abroad.

2 Literature Review

Over the years, translation, as a discipline, has gone through linguistic, cultural, professional, and technological turns. This, subsequently, has shifted the focus of several studies from the conceptualization of translation to investigating translators' knowledge, skills, and expertise, or what is known as TC. Despite the controversial nature of TC, along with its various names, definitions, and descriptions given by translation scholars (Wills, 1996; Bell, 1991; Kiraly, 1995; Malmkjær, 2009; Yang & Li, 2021), there is a scholarly consensus that TC is not a singular entity but comprises a complex set of interrelated sub-components necessary for translators to produce successful translations. Especially as of the 1990s, some theoretical studies (e.g., Kiraly, 2006; 2015) and empirical research projects (e.g., PACTE 2000; 2003; 2017) have been carried out about the definition and modelling of TC to reflect the demands of the translation industry in terms of the skills and competencies required of translators. On the pedagogical side, expert groups (e.g., the European Master's in Translation ‘EMT’) has been initiated to bring academia and industry closer by maintaining the alignment of the curriculum content of translator training programmes with the needs of the translation market (EMT, 2017). It is not the intention here to give a historical background about TC or to discuss the existing models. Yet, it is worth shedding some light on the TC model proposed by the EMT group due to its implementation and usefulness in the current study context. This model was applied in this study because of its technology-orientedness and comprehensive coverage of various aspects of translation technologies as noted by Svoboda & Sosoni 2023. Moreover, the model highlights two educational objectives, i.e., professionalization and employability, which are core principles in the Saudi Vision 2030 and the researchers' broad project.

2.1 TC Model Framework Proposed by EMT

The EMT's model of TC was first established in 2009 on the basis of relevant models to the European context by the Directorate-General for Translation (DGT), which

works in partnership with over 60 MA translator training programmes (EMT Expert Group, 2009). They provided the model as a quality reference for university training programmes to standardize the quality of teaching the competence domains, skills, and professional aspects to be required of professional translators in the translation market. This model was then redrafted in 2017 to live up to the reality of the contemporary translation market, producing five principal competence areas: Language and Culture, Translation, Technology, Personal and Interpersonal, and Service Provision. In their latest version produced in 2022, the EMT board decided to make a minor update to the model framework “to reflect the priorities of European translation programmes, as they prepare graduates for a dynamic and highly technologized workplace” (EMT Competence Framework, 2022:2). By reviewing the TC models succeeding the 2000s, it has been clear that technological competence is considered an indispensable competence area in all the proposed models. Regardless of the different names given to this competence (e.g., ‘technological’, ‘instrumental’ amongst others), these terms “would all refer to one’s knowledge of tools required for the profession” (Oraki & Tajvidi, 2020). Technological competence comprises a set of interdependent sub-competences, which determines what knowings that are supposed to be reflected in translators’ abilities. To acquire this competence area, several studies have stressed that university training programmes should integrate translation technology courses into their study content to produce market-ready translators who are skilled in using translation tools and technologies (e.g., Kenny, 2019).

Bearing in mind that translation jobs are only available for Saudi translators now, who are most likely to be graduates of local universities, several studies have criticized the way translation technologies were being taught in the language-related programmes at Saudi universities pertaining to the rigid curriculum content (Alenezi, 2015), the conventional teaching practices (Omar et al., 2020), and the poor infrastructure of language labs (Abu-ghararah, 2015; Alrumaih, 2021). Such criticisms may raise alarming concerns about the status of the technological competence of professional translators in the Saudi translation market. The next section presents a brief review of related survey-based works considered prominent landmarks in the current study's context.

2.2 Related-Market Surveys

One of the initial research attempts to provide an overview of the translators’ working practices in the Saudi translation market was conducted by Fatani (2009). The findings of this study are significant in providing valuable insights into the translation market’s trends (e.g., globalization), challenges (e.g., translators’ deficiencies), and opportunities (e.g., potential for expansion). This was supported by Abu-ghararah (2017), who warns about the existing gap between the outcomes of academic translation programmes at Saudi universities and the requirements of the translation market. The authors of the two studies anticipate potential growth in the local translation market due to the country’s openness and presence internationally. Although the two studies did not empirically investigate the translators’ technological competence, they reveal a lack of familiarity among professional translators with the use of translation technologies in their daily work; as Fatani (2009: online) described it, the Saudi translation market is

aware of “the deficiencies of Saudi translators who were graduates of local translation and language programs”. Moreover, both studies have mentioned that such competence and quality-related challenges would increase the difficulty for local and global translation companies to recruit a suitable workforce for translation projects.

Another important study was conducted by Alshaikhi et al., (2018) to explore the competence level of professional translators in the Saudi translation market. Seventy-three participants filled out the questionnaire and assessed their level of proficiency in the full set of EMT competence domains, language, translation, technological, intercultural, thematic, information mining, and project management. In terms of the technological competence domain, seven criteria were included to explore how competent the respondents are in performing the IT-related skills. Based on average responses, 41% of the respondents described themselves as proficient in dealing with technology-related skills, whereas 34% were not confident when using technology in their work, rating themselves as sub-competent. The results also show that 21% of the respondents rated their competence as mediocre, indicating that they have modest abilities in performing the technological processes of translation. The study raises alarming concerns about those who confirmed their deficiencies in using translation technologies and calls for monitoring the competence level of professional translators over some period.

Two recent studies are worth mentioning here investigating the Saudi translation market in the post-Covid-19 era. The first study is conducted by Alkhatani (2021) that revealed significant shifts in the working norms of professional translators in the Saudi translation market, particularly towards more remote service provision and digital solutions. It was also found that newcomer translators have much interest in learning and staying up to date with the evolving translation technologies and tools. The author linked this thirst for knowledge to the lack of integration of translation technologies into the curricula of Saudi academic programmes. Similarly, Salamh (2022) focused on the divergence between what is being taught in translator training programmes and the job advertisements in Saudi Arabia pertaining to technology-related skills required of translators. The findings of study highlight that although translation employers prioritize the familiarity with translation digital tools when recruiting new candidates, most professional translators were not confident with their technological competence and felt unequipped with the important skills to use translation technologies efficiently. From these few indications, it has become necessary to conduct this survey-based study to fill a gap in the literature by responding to the following research questions:

1. How do professional translators perceive the teaching status of translation technologies i.e., CAT tools and MT, in BA academic programmes in Saudi Arabia?
2. What is the technological competence status of professional translators in the translation market in Saudi Arabia, and how has it been evolving in the past seven years?

3 Research Method

Using a self-reporting questionnaire, this study collected quantitative data from professional translators in the Saudi translation market. The questionnaire consists of three main sections. The first section collects information about the backgrounds of the participants. The second section seeks to identify the participants' perceptions about the translation technology training, i.e., MT and CAT tools they received during their BA studies. The third section contains thirteen criteria related to translation technologies, which were adapted from the OPTIMALE survey (Toudic, 2012). Here, participants were asked to assess their level of competence using a Likert-type scale ranging from excellent, competent, average, sub-competent, and weak. The participants were also provided with a not applicable (N/A) option if they did not require the skill in their translation work. The survey was prepared using the Qualtrics software tool and circulated online from December 2021 to February 2022. In total, 404 responses were returned, 248 of which were completed and valid for analysis. Compared to similar studies conducted in Arab translation markets, the study's sample size and response rate are encouraging. Notably, the reliability of the Likert-scale section was measured by calculating the Cronbach alpha value of 0.892, denoting a high degree of reliability. In terms of ethical considerations, this study has gained ethical approval and met all the requirements indicated by the ethics committee of Swansea University.

4 Results and Discussion

Based on the sheer use of descriptive statistical analysis, this section first illustrates some information about the participants' demographics in a general manner, followed by their answers to the questions in sections two and three of the survey. The results show a participation predominance among females 78% compared to males 22%. The bulk of participants 86% are below 35 years old, denoting that they are from the rookie workforce of translators in the Saudi translation market. This result was also supported by the respondents' length of experience since 72% have less than five years of experience as translators, which indicates that they are fresh translators. In addition, the results show that the responses came from various kinds of sectors, including private (38%), semi-governmental (11%) and governmental (24%) organizations, as well as LSPs (27%). Regarding their academic background, 70% of the respondents completed the BA studies whereas the remaining 30% hold postgraduate degrees either MA or PhD. Moreover, the participants were asked about their study backgrounds and the type of BA programme they graduated from. Although 50% of the respondents are graduates of translator training programmes, the results indicate that graduates of other BA study backgrounds also occupy the translation professions in the Saudi translation market, representing as follows; 40% came from English learning programmes, 7% from linguistics or literature programmes and a marginal percentage 3% studied in other academic programmes such as Arabic language or Islamic studies. This result flags the concern about the technological competence level of translators who graduated from non-translation programmes. According to Alsolami (2022), although half of the

English language programmes at Saudi universities state an objective to prepare professional translators for the market, only three (out of eight) integrate the teaching of translation technologies into their curriculum, indicating a potential lack of technological competence among would-be translators. The teaching content and practices of the English learning programmes might not be sufficient to meet the reality and requirements of the modern translation market. This finding should encourage significant reforms in the curriculums of these programmes to include translation technology training to ensure providing the translation market with translators with a sufficient level of technological competence.

The participants were also required to identify the kind of teaching used to teach MT and CAT tools in the academic programmes they graduated from. Figure 2 below displays the results of the teaching status of MT and CAT tools, as indicated by the respondents.

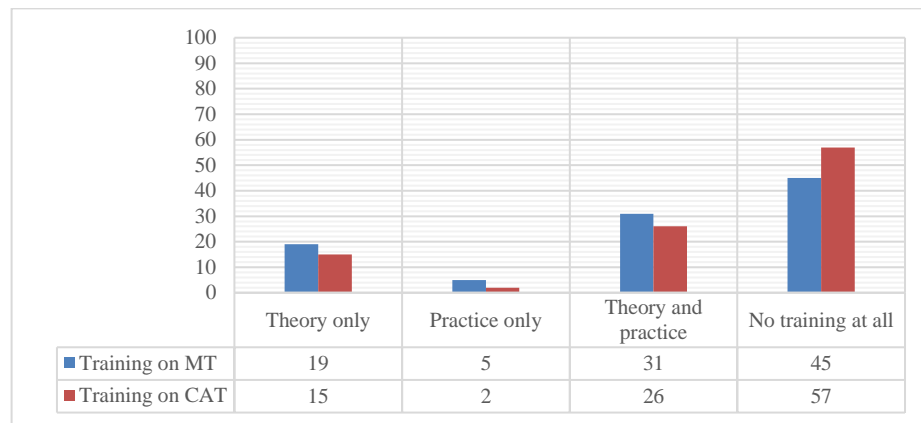


Fig. 1. Teaching status of CAT tools and MT (respondents' views).

The above figure shows that about half of the respondents did not receive any training on MT, in addition to roughly one-fifth who learnt MT theoretically only without any practical training. This makes up almost two-thirds of the respondents who were not exposed to MT training during their BA studies. This finding is consistent with Abu-ghararah's (2015) study that investigated the availability of technology and language resources in translation programmes at Saudi universities. It was found that most students 80% confirmed the absence of MT teaching in these academic programmes; thus, they lack technological competence. Regarding their perceptions towards CAT tools training, most respondents described the situation as worse than training on MT. The results show that more than half of the respondents were not trained at all in using CAT tools during their BA studies, in addition to 15% who were only provided with theoretical knowledge about CAT tools. This increased the number of translators whose abilities are more likely to be deficient in using CAT tools when they start their translation careers. The results indicates that translation technology training has not adequately been provided for trainee translators in Saudi academic programmes since almost two-thirds of the respondents were not trained properly to use MT or CAT tools during their

BA studies. This finding supports the previous studies such as Abu-ghararah (2017) and Alrumaih (2021), who pointed out that the gap between translator training programmes and the translation industry in Saudi Arabia is still perceived, particularly concerning the insufficient integration of translation technologies into the curriculum content taught in the academic programmes. Specifically, Alrumaih (2021) reports that there is a diversity in the way of delivering translation technology courses, whether practically or theoretically, to students in the translation programmes at Riyadh-based universities. The author found that these academic programmes lack the provision of practical training on CAT tools for trainee translators. There may be a necessity for policymakers and curriculum designers to rethink about the teaching content and methods adopted to teach translation technology courses in Saudi academic programmes.

Moreover, the respondents were asked to rate their technological competence in thirteen statements categorized into four skill domains. These domains are as follows: skills in MT (3 criteria), skills in CAT tools (4 criteria), skills in localization (2 criteria) and Skills in other technology-related activities (4 criteria). It is worth mentioning that these domains are increasingly recognized in the modern translation industry as technology language trends, as revealed in the Nimdzi Language Technology Atlas published in 2022.

Technically speaking, the results of this section are reported and discussed based on the average percentages of the responses to provide estimates of the translators' level of competence in each technological domain. Figure 3 below provides relevant details to the participants' three abilities to pre-edit texts for MT, post-edit MT, and configure MT systems.

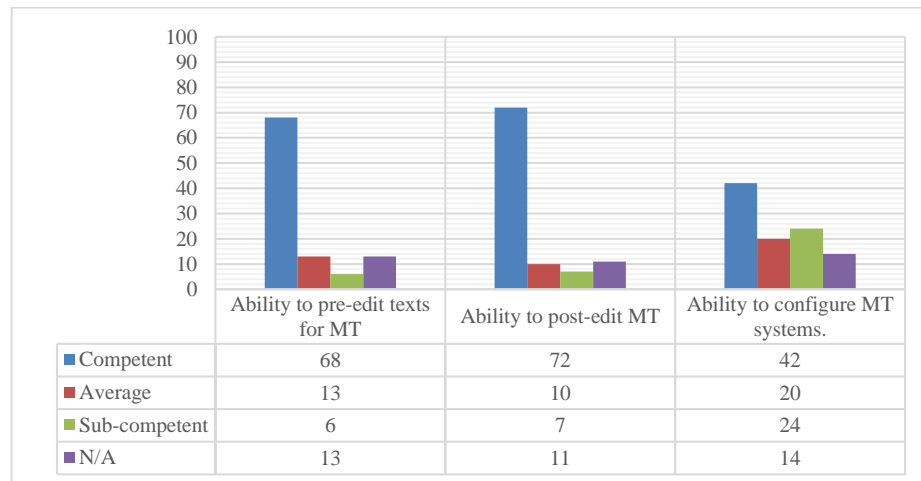


Fig. 2. Translators' proficiency in MT-related skills.

The results indicate that MT is widely accepted as part of the normal translation work in the modern translation market in Saudi Arabia since only an average of 13% of the respondents who said that MT is unnecessary for their work. This result aligns with what Alkhatani (2021) reported about the increasing adoption of MT among translators

in the Saudi market. The overall assessment shows that an average of 61% of the respondents describe their competence status as proficient in MT-related skills. This should be seen as a positive development in the Saudi translation market since this technology (MT) is becoming increasingly prevalent in the translation market worldwide. However, the fact that an average of 12% of the respondents labelled themselves incompetent indicates that they are not confident about their skills in using MT. Moreover, the results show that an average of 14% of the respondents confirmed their moderate abilities in MT-related skills, which increases the number to be just over one-quarter of the respondents with a limited level of proficiency in dealing with MT-related skills. The latest version of the EMT model has enhanced the crucial incorporation of the skill of interacting with MT within the overall professional translation competence (EMT, 2017). According to the recent Nimdzi market survey (2023), the majority of the surveyed translation companies have witnessed a significantly greater use of MT during the second quarter of 2023 compared to the preceding quarter, up to 60.4% from 42.9%. Moreover, Salamah (2022) mentioned that many job announcements posted by translation employers in the Saudi translation market require proficient candidates in MT. Having the ‘Saudization’ process in mind, translators with a low level of competence or completely deficient to use MT might find difficulties to find a job in the competitive translation market. The results indicate that the MT domain might be an area of strength for most professional translators; yet there is still a need for improvements in terms of providing additional training for translation professionals to develop their competence in using MT more efficiently.

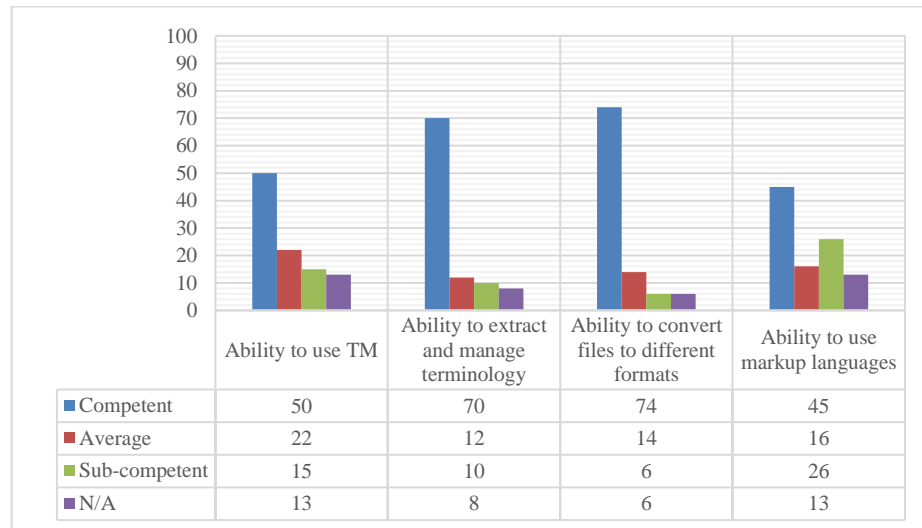


Fig. 3. Translators’ proficiency in CAT-related skills.

Like the use of MT, the overall assessment shows that CAT tools are widely used among most professional translators in the Saudi translation market since only an average of 10% stated that this technology is not required in their work. Perhaps, translators

are less likely to use CAT tools if they work in translation fields where using CAT tools is not common (e.g., literary translation field). The results illustrate that an average of 60% of the respondents rated their abilities as competent in the four CAT tools-related skills. This could be a positive sign of the status of the translators' level of proficiency in adopting CAT tools in their work. However, an average of 14% of the respondents described themselves as deficient in dealing with CAT tools, in addition to an average of 16% who have average abilities in this area. Such a deficiency in skills and knowledge of CAT tools could influence the translator's productivity and work quality, which, as a result, have potential impacts on both the reputation of translators and the quality of the translation market. In their study, Alshaikhi et al., (2018) reported similar findings about the translators' lack of familiarity with the activities related to translation memory (TM) and termbase (TB) during their work. In the same regard, Alkhatani (2021) also stressed the essential demand to equip professional translators with the necessary skills to use the advanced features and customization of CAT tools. Therefore, more comprehensive training sessions are needed to ensure that translation professionals are fully equipped to meet the needs of the modern translation market.

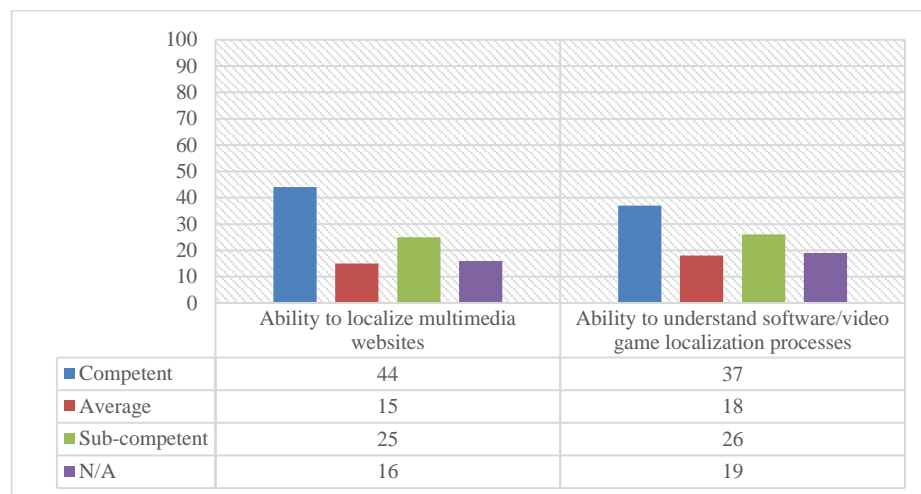


Fig. 4. Translators' proficiency in localization-related skills.

The above figure displays the responses to the two items related to the localization domain, which registered similar ratings by the respondents. The results show that an average of 40% of the respondents rated themselves as competent in performing the localization processes (e.g., multimedia websites). However, a significant percentage of the respondents rated their competence level as only moderately competent 16% or completely deficient 26%, making up a total average percentage of 42% of the respondents without the necessary skills and knowledge to deal with the complexities of localization processes. Unlike traditional text-based translation, localization projects need a high level of proficiency from translators to successfully produce high-quality localized content. The results indicate a lack of competence among professional translators in

localization-related skills, which potentially impacts the quality and efficiency of their work. In other words, a large percentage of professional translators felt unprepared to provide localization services or involve in large-scale projects that need translators with a sufficient level of competence in the localization domain. Therefore, it could be stated that this skill domain is an area needing more improvements for the majority of the respondents. Previous research has not adequately explored the localization aspect of the Saudi translation market and the competence level of translators (localizers), which makes it difficult to monitor the translators' level of competence in this specific domain. However, some resources in the literature pointed out that the requirements of translation employers have been long misaligned with the level of competence of job seekers (Fatani, 2009; Abu-ghararah, 2017; Salamah, 2021).

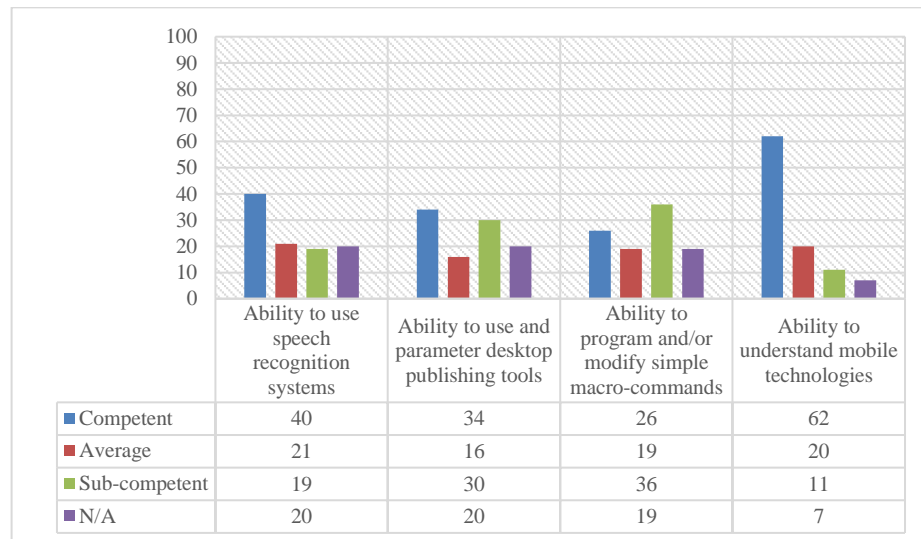


Fig. 5. Translators' proficiency in different technology-related skills.

Similar to the ratings given in the localization domain, the overall assessment shows that an average of 40% of the respondents rated their level of proficiency as competent in the four skills shown in Figure 5. However, a relatively high percentage of respondents rated themselves as either moderate 19% or sub-competent 24%, which is a total of 43% on average. This result indicates a lack of understanding and proficiency among the majority of participants to implement more specialized technologies, which has resulted in translators feeling unconfident in dealing with speech recognition systems, desktop publishing tools, programming macro-commands and mobile technologies. In a growing translation market, translators should not be lagged behind evolving technologies to avoid being at a disadvantage in competing for work. The results suggest that acquiring new skills related to various types of technologies and improving them are necessary to facilitate the translation process for translators and keep them up to date with the changing needs of the modern translation market in the 21st century.

5 Conclusion

Having surveyed an important stakeholder in the translation market, i.e., professional translators, the current paper is a call for academics to think of new ways of teaching translation technologies to match the requirements of the modern translation industry. Despite the technological ill-preparedness reported by the respondents during their BA academic studies, the self-assessment survey indicates a relatively high competence level among more than half of the respondents in interacting with MT and dealing with CAT tools in the Saudi translation market. However, the results also report that there are technological competence areas in which translators need further training and support to adapt to the changing technological landscape. The participants show some deficiencies in dealing with the processes of localization and using more advanced technologies such as programming macro commands. Moreover, the paper suggests that translation scholars and researchers may need to revisit the concept of technological competence and what that actually means in light of recent technological developments and how the various tasks of the translation process will be divided between humans and machines in a technologically driven industry. Translators' approaches to technological competence should also be reviewed to encompass the ability to adapt to new technologies and interact with the machines effectively.

References

1. Abu-ghararah, B.: The availability and use of technology and learning resources in translation programmes. *Arab World English Journal (AWEJ) Special Issue on CALL*, (2015).
2. Abu-ghararah, B.: The gap between translator training and the translation industry in Saudi Arabia. *AWEJ for translation & Literary Studies Volume 1*, (2017).
3. Alenezi, A.: Development of translation curricula at undergraduate translation courses in Saudi universities: exploring student needs and market demands. PhD Thesis, University of Leicester, (2015).
4. Alkhatnai, M.: Perceptions, Skills, and Technologies for the Revitalization of Translation Industry in the Post COVID-19 Era: An Empirical Evidence from Saudi Arabia. *Journal of Foreign Language Teaching and Translation Studies*, 6(3), 71-96 (2021).
5. Al-Rumaih, L. A.: The integration of computer-aided translation tools in translator-training programs in Saudi Universities: Toward a more visible state. *AWEJ for Translation & Literary Studies*, 5(1) (2021).
6. Alshaikhi, T. Rothwell, A. & Parra, M. F.: Two sides of the coin: Mapping translators' competence criteria onto the market requirements: The case of Saudi Arabia. In E. Poirier and D. Gallego-Hernandez (eds). *Business and Institutional Translation: New Insights and Reflections*, 203-218 (2018).
7. Alsolami, A.: Mapping Translation Technology Provision in Saudi University Degrees: Two Different Pathways. In Castilho, S., Caro Quintana, R., Stasimioti, M., and Sosoni, V. 2023. *Proceedings of the New Trends in Translation and Technology Conference*. 4-6 July 2022, Rhodes Island, Greece.
8. Bell, R, T.: *Translation and translating: Theory and practice*. Vol. 298. London: Longman, 1991.

9. Bowker, L., Pastor, G. C.: Translation Technology. In: *The Oxford Handbook of Computational Linguistics*. Oxford University Press (2015). <https://doi.org/10.1093/oxfordhb/9780199573691.013.007>, accessed 15 June 2023.
10. EMT Expert Group.: *Competences for professional translators, experts in multilingual and multimedia communication* (2009).
11. EMT Board.: *EMT Competence Framework-2017*, https://commission.europa.eu/system/files/2018-02/emt_competence_fw_2017_en_web.pdf, last accessed 2023/06/20
12. *EMT Competence Framework-2022*, https://commission.europa.eu/system/files/2022-11/emt_competence_fw_2022_en.pdf, last accessed 2023/06/20
13. Fatani, A.: The state of the translation industry in Saudi Arabia. *Translation Journal*, 13(4), 1-8 (2009).
14. Kenny, D.: Technology and translator training. *The Routledge Handbook of Translation and Technology*, 498–515 (2019). <https://doi.org/10.4324/9781315311258-30>
15. Kiraly, D. C.: *Pathways to translation: Pedagogy and process*, Kent (Ohio), the Kent State University Press. *Meta: Journal Des Traducteurs* 41(1), 176 (1995). <https://doi.org/10.7202/002159ar>
16. Kiraly, D. C.: Beyond social constructivism: Complexity theory and translator education. *Translation and Interpreting Studies* 1(1), 68–86 (2006). <https://doi.org/10.1075/tis.1.1.05kir>
17. Kiraly, D. C.: Occasioning translator competence: moving beyond social constructivism toward a postmodern alternative to instructionism. *Translation and Interpreting Studies Vol. 10*(1), 8-32 (2015). <https://doi.org/10.1075/tis.10.1.02kir>
18. Malmkjær, K.: What is translation competence?. *Revue française de linguistique appliquée*, 14(1), 121-134 (2009).
19. Nimdzi. *Machine Translation: Types of Services on Offer and Usage by LSPs*, <https://www.nimdzi.com/machine-translation-types-of-mt-services-offered-and-average-usage-by-lsps/>, last accessed 2023/06/20
20. Nimdzi. *The Nimdzi Language Technology Atlas*, (2022), <https://www.nimdzi.com/language-technology-atlas/#language-technology-trends>, last accessed 2023/06/20
21. O'Hagan, M.: Introduction: Translation and technology: disruptive entanglement of human and machine. In *The Routledge handbook of translation and technology* 1-18 (2019).
22. Olohan, M.: Translators and translation technology: The dance of agency. *Translation studies*, 4(3), 342-357 (2011).
23. Omar, A., F., A., El-Nabawi, I.: The impact of translation software on improving the performance of translation majors. *International Journal of Advanced Computer Science and Applications*, 11(4) (2020). <https://doi.org/10.14569/ijacsa.2020.0110439>
24. Oraki, A., & Tajvidi, G.: Training translators and interpreters: The need for a competence-based approach in designing university curricula. *Iranian Journal of English for Academic Purposes*, 9(2), 42-56 (2020).
25. PACTE Group.: *Acquiring translation competence: hypotheses and methodological problems of a research project*, In the 4th International Congress on Translation, pp. 99-106. John Benjamins, Barcelona (2000) <https://doi.org/10.1075/btl.32.13pac>.
26. PACTE Group.: *Building a translation competence model*, In Alves, F. (Ed.), *Triangulating Translation: Perspectives in Process Oriented Research*, John Benjamins, Amsterdam/ Philadelphia, 43-66 (2003) <https://doi.org/10.1075/btl.45.06pac>.
27. PACTE Group.: *PACTE Translation Competence Model. Researching Translation Competence by PACTE Group*, 35–42 (2017). <https://doi.org/10.1075/btl.127.02pac>

28. Salamah, D.: Translation competence and the translation job market in Saudi Arabia: Investigating recruitment practices and job-market readiness. *Saudi Journal of Language Studies*, 2(4), 236–258 (2022). <https://doi.org/10.1108/sjls-08-2022-0064>
29. Svoboda, T., & Sosoni, V.: Institutional translator training in language and translation technologies. In *Institutional Translator Training* pp. 73-91 (2023). Routledge.
30. Toudic, D.: The OPTIMALE employer survey and consultation (2012). Retrieved from https://euatc.org/images/Optimale_WP4_Synthesis_report.pdf
31. Wilss, W.: Knowledge and skills in translator behavior. *Knowledge and Skills in Translator Behavior*, 1-273 (1996).
32. Yang, Z., & Li, D.: Translation competence revisited: Toward a pedagogical model of translation competence. *Advances in Cognitive Translation Studies*, 109–138 (2021). https://doi.org/10.1007/978-981-16-2070-6_6.

An Exploration of Risk in the Use of MT in Healthcare Settings with Abbreviations as a Use Case

Eleanor Taylor-Stilgoe, Constantin Orăsan^[0000-0003-2067-8890] and
Félix do Carmo^[0000-0003-4193-3854]

University of Surrey, Guildford, Surrey, GU2 7XH, United Kingdom
{e.j.taylor-stilgoe,c.orasan,f.docarmo}@surrey.ac.uk

Abstract. When faced with language barriers, UK healthcare staff have found themselves turning to machine translation (MT) – predominantly Google Translate – to fulfil their duty of care to patients [3, 16, 23]. Despite the risks potentially posed by the use of MT in such complex and sensitive situations, little research currently exists as to healthcare staff awareness of these risks in real-life settings. This gap is particularly notable concerning the use of MT with patient medical record information compared with interpersonal situations and patient-oriented documentation [6, 7, 8, 14, 16, 22, 27]. While research has been conducted into the perceptions and practices of the general population concerning MT use in largely lower-stakes contexts [31], research on the extent to which these transfer to higher-stakes settings remains lacking. The contribution this paper aims to make is therefore twofold: to investigate the impact of MT on patient medical record documentation and to explore the extent to which healthcare staff are aware of the risks potentially posed by its use. In this paper, we selected contextualised medical abbreviation examples from authoritative French and Spanish clinical corpora [9, 15] to serve as a use case, abbreviations having previously been shown to pose an increased risk for patient harm even prior to their translation with Google Translate [2, 5, 12, 19, 25, 28]. Examples containing higher-risk MT errors were presented to healthcare staff to ascertain their perceptions and risk awareness as part of semi-structured interviews. Whilst these interviews remain ongoing, this paper presents the findings on risks identified in the use of MT with patient medical documentation, and the responses obtained thus far.

Keywords: machine translation, Google Translate, healthcare, patient medical records, medical abbreviations, risk awareness

1 Introduction

The National Health Service (NHS) in England formally cautions its staff against using online MT services on the grounds that ‘*there is no assurance of the quality of the translations*’ [17]. Despite this, cases abound in which healthcare staff have found themselves resorting to non-domain specific, commercially available MT, such as Google Translate (GT), when providing interpersonal or written assistance to patients with limited to no English language proficiency [3, 16, 23]. Vieira et al. [30] note that ‘*research on the implications of the widespread and potentially uninformed use of this technology remains sparse*’. Further to this, despite patient medical records constituting an arguably central component of patient care, there is currently next to no literature on the implications of their

use with MT. The need for further research in this specific yet significant subdomain is underscored by the comparative attention paid to the use of MT as a form of interpreting aid in interpersonal situations and for patient-oriented documentation [6, 7, 8, 14, 16, 22, 27]. The contribution this paper therefore aims to provide is an investigation of the impact of MT on patient medical documentation, and an examination of the extent to which healthcare staff are aware of the risks potentially posed by its use.

To this end, contextualised examples containing medical abbreviations were selected as a use case, these ubiquitous features of the healthcare domain having been shown to pose an increased risk for patient harm even at the monolingual level [2, 5, 12, 19, 25, 28]. This paper presents an analysis of the impact of GT on specific abbreviation examples in terms of the errors produced at a linguistic level and the potential risks to patient safety arising from its use. Examples containing higher-risk MT errors were presented to healthcare staff within the context of semi-structured interviews to ascertain their perceptions of and risk awareness of the same. While interviews with healthcare staff remain ongoing, the preliminary conclusions drawn thus far are provided in Section 8 below.

2 Related Work

The misinterpretation and/or misuse of medical abbreviations in healthcare settings has been identified in the literature as a significant contributing factor towards increased potential for patient harm, even at the monolingual level [2, 5, 12, 19, 25, 28], though scant research comparatively exists with regard to their translation using MT. It is for precisely this reason that abbreviations were selected as a use case with which to analyse phenomena for which the risk to patient safety in such higher-stakes contexts may potentially be further exacerbated by the use of commercially available and non-domain specific MT.

In an attempt to overcome the difficulties posed by language barriers in healthcare settings, various domain-specific applications have been developed over the last twenty years with a view to improving delivery of care to patients from diverse cultural and linguistic backgrounds [1, 4, 18, 26]. Further to this, a range of mobile translation applications developed specifically for providing communication support to both staff and patients unable to speak the dominant language of the healthcare setting have been subject to formal evaluation as to their effectiveness in practice [11, 20, 21, 24, 29]. Despite these developments, commercial MT in general and Google Translate in particular remain prevalent in both academic and grey literature discussing the use of technologies in healthcare settings [3, 6, 7, 8, 14, 16, 22, 27]. The ready availability, convenience of use and cost-effectiveness of MT appear to make it a 'better than nothing' alternative for staff who must balance their duty of care against time, cost, and/or resourcing constraints [3, 16, 23]. As noted by Vieira et al. [30], *'language barriers coupled with funding pressures and other practical difficulties expose doctors and patients to a dilemma where MT, albeit risky, is perceived as the easiest route to cross-linguistic communication'*. Google Translate was therefore selected for use in the present research, reflecting as this does its documented use in real-life healthcare contexts.

Furthermore, while research has been conducted into the perceptions and practices of the general population concerning MT use in primarily lower-stakes contexts [31], studies on the extent to which these transfer to such higher-stakes settings as healthcare remain limited. This paper therefore also seeks to contribute to the knowledge gap in regard to the extent to

which healthcare staff are aware of the risks potentially posed by its use when applied to professional settings involving patient care.

Finally, the existing literature on the use of MT in healthcare settings is largely focused on its use as a form of interpreting aid in interpersonal situations and for patient-oriented documentation [6, 7, 8, 14, 16, 22, 27]. As such, this paper aims to examine its use within the context of patient medical record information, this being a particularly understudied yet significant aspect of an already understudied area.

3 Data Research Methodology

3.1 Data Collection

Medical abbreviations were selected as a use case with which to investigate and analyse the creation or amplification of risk through their use with MT. This decision was made on account of such abbreviations having previously been identified as posing an increased risk for patient harm even prior to translation [2, 5, 12, 19, 25, 28]. It should be noted that ‘abbreviation’ is, per Sheppard et al. [25], used as an umbrella term for ‘*any shortened form of a word [...] any acronym [...] any contracture [...] and any initialism*’.

It was first necessary to select at least one in-domain, non-English SL corpus with which to identify and isolate instances of relevant phenomena. A key criterion was the prioritisation of reliable clinical data from authoritative sources. The aim of this was to echo as authentically as possible the medical nature and linguistic features of written information encountered in healthcare environments, allowing for a more realistic assessment of the potential risks involved in its translation with MT. The decision was made to begin with source languages with which the first author was already familiar, namely French and Spanish.

The first corpus selected was the CAS corpus, a French clinical corpus comprising clinical cases published in scientific literature and both legal and medical training resources [9]. This contained a mixture of real, de-identified and fake patients used for medical educational purposes and comprised 4,900 clinical cases. The second corpus selected was CodiEsp, a Spanish clinical corpus comprising a randomly sampled development dataset (250 cases), test set (250 cases) and training set (500 cases), along with an unannotated background set containing a further 2,751 cases [15]. A notable advantage offered by both corpora was their coverage of various specialties, patient comorbidities and case histories written at the expert-to-expert level, as would be expected from the type of real-life healthcare documentation on which both were based.

3.2 Data Preparation and Processing

Given the size of the corpora used, a simple custom desktop application with which to identify and isolate relevant data instances was designed by the first author and implemented by a professional software developer using Microsoft Visual Studio and C#. This allowed for the filtering of sentences from the corpus according to manually configurable and selectable search patterns, implemented using regular expressions. The option to manually add matches containing irrelevant instances to a list of exclusions was also incorporated, as was a feature enabling exportation of selected matches to an Excel file for use with MT and analysis in

Sketch Engine [13]. Any noise present in the data, such as spelling or grammatical errors, was preserved with the aim of recreating in the MT output those issues encountered in real-life settings. The paragraph within which each instance occurred was also preserved according to the preceding and subsequent line breaks to ensure that sufficient context was retained. The retrieved matches were then translated with GT.

4 Data Examples: Experimentation, Results, and Analysis

A random set of SL matches were analysed to explore any issues arising from their translation with GT. In addition to focusing on abbreviations, any other relevant phenomena were acknowledged and recorded for reference, with the research remaining focused on the use case. The following examples were considered particularly notable in terms of their potential for increased risk to patient safety and were thus selected for use in interviews with healthcare staff. As follows (emphasis and/or [sic] added for clarity):

Example 1: ‘AO’

ST: En el examen inicial presentaba una AV con su corrección (AV CC) de 0,2 en AO que mejoraba a 0,4 con estenopeico. La motilidad pupilar era normal, la presión intraocular (PIO) de 18mm Hg en AO y no presentaba alteraciones en la biomicroscopia anterior (BMA). La refracción bajo cicloplejía (RBC) era de -4,50 en AO. En el fondo de ojo (FO) se apreciaban estrías retinianas en la mácula de AO. La tomografía axial computarizada (TAC) no reveló anomalías en la región orbitaria ni asimetrías ni alteraciones en la posición del diafragma iridocristaliniano.

GT: In the initial examination, she presented a VA corrected for her (VA CC) of 0.2 in AO that improved to 0.4 with pinhole. Pupillary motility was normal, intraocular pressure (IOP) was 18mmHg in AO, and she had no alterations in anterior biomicroscopy (AMB). Refraction under cycloplegia (RBC) was -4.50 in AO. Retinal striae were seen in the eye fundus (FO) in the AO macula. Computerized axial tomography (CAT) revealed no abnormalities in the orbital region or asymmetries or alterations in the position of the iridocrystalline diaphragm.

This Spanish to English ophthalmological example contains four instances in which the abbreviation ‘AO’ (‘ambos ojos’ in Spanish, for which the correction translation would be the Latin ‘OU’, ‘oculus uterque’ or English ‘BE’, ‘both eyes’) remains wholly untranslated in the MT, seemingly due to the lack of a full term preceding the SL abbreviation. The result is the omission of central clinical information as to the relevant examination having been conducted on both eyes. While other abbreviations present in the MT, such as ‘RBC’ (‘refracción bajo cicloplejía’/‘cycloplegic refraction’, ‘CRx’) and ‘FO’ (‘fondo de ojo’/[eye] fundus’) remain either mistranslated or untranslated, these are at least preceded by the full if only broadly rendered terms, whereas this surrounding context remains absent for ‘AO’.

Example 2: ‘DRS’

ST: Au cours des dix dernières années, elle a subi quatre coronarographies, neuf échographies cardiaques et douze épreuves d’effort au tapis roulant. Depuis la dernière dilatation, quatre ans auparavant, aucune lésion significative ou donnée probante de resténose ne sont notées aux

*coronarographies. Les pressions pulmonaires sont normales. Aucune maladie valvulaire n'apparaît aux échographies. Les paramètres électrophysiologiques en relation avec les suivis liés à son stimulateur cardiaque défibrillateur sont adéquats[sic]; aucune anomalie significative n'apparaît et aucune thérapie ne lui est donc appliquée. Bien que cette dame ne soit pas suivie dans cet établissement, elle nous assure que son lupus est stable. D'ailleurs, la protéine C réactive est normale, la tension artérielle est bien contrôlée, et les douleurs angineuses n'ont pas un caractère péricardique. Enfin, elle n'est connue pour aucun antécédent psychiatrique. Pourtant, la patiente continue de présenter des **DRS**.*

*GT: Over the past ten years, she has had four coronary angiograms, nine cardiac ultrasounds and twelve treadmill stress tests. Since the last dilation, four years earlier, no significant lesions or evidence of restenosis have been noted on coronary angiograms. Pulmonary pressures are normal. No valvular disease appears on ultrasound. The electrophysiological parameters in relation to the follow-ups related to his defibrillator pacemaker are adequate; no significant abnormality appears and no therapy is therefore applied to him. Although this lady is not followed in this establishment, she assures us that her lupus is stable. Moreover, C-reactive protein is normal, blood pressure is well controlled, and angina pain is not pericardial in nature. Finally, she is not known to have any psychiatric history. However, the patient continues to present with **DRS**.*

This largely cardiological French to English example, which shifts to psychiatry in the penultimate sentence, reflects a complete failure by the MT to translate 'DRS' ('douleur(s) rétrosternale(s)'), for which the correct translation in English would be 'retrosternal pain'. While this non-translation in the MT also appears to have arisen from the lack of a full term preceding the SL abbreviation, by comparison with 'AO' for which a directly equivalent medical abbreviation does indeed exist in either Latin or English, both options being in use in modern healthcare settings, no such direct equivalent exists in English for 'retrosternal pain'. Despite being the only abbreviation in the paragraph, 'DRS' is of central importance to identifying the patient's symptoms, described as these are as ongoing. The ambiguity arising from this non-translation may hinder the patient's treatment as a result.

In view of the untranslated abbreviations in the MTs and the apparent centrality of these to comprehension, it was speculated that the ambiguity arising from such issues may at best hinder effective care delivery or, at worst, actively increase the potential for patient harm in real-life healthcare settings. Interviews with healthcare staff were planned to explore their views on this potential for increased risk accordingly.

5 Interview Methodology

5.1 Recruitment Process

Participants approached for recruitment included both medically qualified professionals and those performing non-clinical healthcare roles, this professional diversity being considered directly relevant to the type of information each might seek to identify in the texts, their respective degrees of clinical risk awareness, and their approaches to risk management. No specific criteria were assigned concerning participants' foreign language knowledge, length of professional service or specific care setting, these aspects all being considered to offer both more realistic diversity in terms of background and potentially interesting analyses.

Recruitment was conducted via snowball sampling. The study was granted a Favourable Ethical Opinion by the University of Surrey Ethical Committee in February 2023.

5.2 Interviews: Data Example Presentation and MT Output Scoring

During the interview process, the selected abbreviation examples were presented consecutively to healthcare staff for their responses. A blind study approach was considered essential to enhancing interview response quality on the basis that, in addition to more closely mimicking the types of situations liable to occur in real life, taking such an approach would assist in safeguarding against priming or increasing the risk of social desirability bias among participants. The full interview questions are listed in the Appendices below.

Participants were first shown the ST on a blind basis, their being neither alerted to the presence of potentially higher-risk phenomena nor given any prior explanation as to the content beyond its pertaining to patient medical documentation and provenance from authoritative clinical data. Furthermore, the ST was framed in terms of a scenario specific to what each participant might encounter in the course of their particular professional role. This contextualisation was considered important to ensuring that participants fully grasped the aim of the exercise beyond the purely linguistic and increased their confidence in being able to provide relevant and comprehensive information based on their approach to and/or experience of such documentation in real-life settings.

Participants were first asked what information they might look to identify in the ST and for what purposes (Q1) to gain an insight into their particular priorities and immediate thought processes (e.g., identification of specialty, degree of clinical urgency, etc.), should these vary by professional role. They were then presented with the corresponding MT in English, retaining access to the ST for reference purposes, and asked what action they would take in the event of identifying a potential error in the example MT output (Q2). Where participants were unable to identify any errors in the given examples, they were permitted to answer the question on a more general basis, again with a view to understanding their thought processes and practices when managing such potential risks in real life.

Participants were then asked to score their degree of confidence in using the MT output in a real-life professional situation, along with the reason for their score (Q3). The following 6-point Likert scale was applied: 0 = 'not at all confident'; 1 = 'slightly confident'; 2 = 'somewhat confident'; 3 = 'fairly confident'; 4 = 'confident'; 5 = 'completely confident'.

This first round of scoring was followed by a discussion with the participant as to the general content of the MT (e.g. the specialty, scenario, and the nature of the information being conveyed), the SL meaning of the centrally problematic abbreviation in question and its corresponding translation in English, and the potential clinical impact of its remaining untranslated in the MT output on care delivery. On being given this information, participants were invited to share their thoughts on these elements. They were also at liberty to comment on any other abbreviations or higher-risk phenomena they had identified during or as a result of this discussion, given the scope this provided for gaining insight into any changes in their perceptions of the MT and/or their degree of risk awareness.

Following the discussion stage, all participants were invited to provide another confidence score and corresponding reason for their score (Q3), even if this remained unchanged. Medically qualified participants were asked to provide a score for the degree of potential risk to patient safety they perceived to be present in the MT and asked to provide a reason for their score (Q4). Only medically qualified participants were invited to respond to Q4 in order

to preserve the clinical authority of the responses. A 4-point Likert scale adapted from Brunetti et al. [5] was applied accordingly: 0 = ‘no capacity to cause harm’; 1 = ‘capacity to cause harm’; 2 = ‘capacity to contribute to or result in temporary harm requiring intervention’; 3 = ‘capacity to contribute to or result in death’.

Finally, all participants were given the opportunity to add any closing comments prior to concluding the interview. This was phrased as an unofficial open question so as to elicit as natural a response as possible from participants.

6 Interviews: Data Example Results

As interviews remain ongoing, only a small sample have been fully analysed and are included in this paper accordingly. These have nonetheless yielded interesting insights in regard to participants’ awareness of the risk present in the examples provided. It should be noted that, due to the depth of analysis applicable to each participant, it was necessary to exclude the results for the ST alone (Q1). The core findings on participants’ responses to the MT output are presented according to the order in which the interview was conducted.

Participant 1 (IV1): Referral Hub Supervisor

IV1, a native European Portuguese speaker, had worked in primarily administrative roles within the NHS, including Community Outpatients and A&E, over almost ten years. While their current and prior roles were not directly clinical in nature, they noted having ‘*some clinical background as I was in the ambulances [...] but my main role was a coordinator in A&E*’, describing this as ‘*admin with the knowledge of clinical*’. During the course of the interview, IV1 stated that they could ‘*understand more or less Spanish*’ but not French. Both texts were presented to IV1 in the context of a patient arriving in A&E with medical information important to their continuity of care.

Example 1: ‘AO’

Q2: IV1 placed particular emphasis on their understanding that, not being medically qualified, ‘*it wouldn’t be my place to make a clinical decision*’ in regard to the MT’s reliability and as such would pass the text on to a clinician. They added that they would escalate it and seek an interpreter or translator ‘*to be sure that this is the right translation*’.

Q3 (confidence scoring prior to discussion): Of particular note was IV1’s mistrust in the MT’s reliability, seemingly owing to their unfamiliarity with the medical terminology used (‘pinhole’), for which a register shift from expert-to-expert to expert-to-lay was instead perceived as having occurred: ‘*[...] it doesn’t sound clinical [or] professional [...] it sounds like the type of language you would be talking to patient for them to understand*’. IV1 otherwise stated that ‘*all the rest, it does seem spot on [...] I’m really very, very impressed [...]*’. A score of four (‘confident’) was assigned, with IV1 noting that the ambiguity surrounding ‘pinhole’ was the main reason for questioning the MT’s fitness for purpose.

Discussion: IV1 indicated that the numerical figure preceding ‘AO’ had led them to believe that this referred to a medication name rather than a body part, thus not initially catching their

attention as higher-risk: *'[...] I thought it was a saline, a sort of medication, I just jumped it ahead'*. They added that, without the preceding full term from which to glean more context for 'AO', *'you wouldn't know [what it was]'* and, in attempting to identify it as a medication, clinicians *'would be wasting [...] hours no end, because they would not find it'*. This discussion appeared to galvanise IV1 into examining other abbreviations present in the MT more closely, with their independently making the observation that 'FO' ('fondo de ojo'/'[eye] fundus') likewise remained untranslated.

Q3 (confidence scoring following discussion): IV1 reduced their score from four ('confident') to three ('fairly confident') on the grounds that *'in this situation [...] ['AO'] is important, but it's not essential'* according to their perception that the clinical picture was normal, though they had interpreted the text as pertaining only to one eye, rather than both.

Example 2: 'DRS'

Q2: The same answer as previously applies to this question.

Q3 (confidence scoring prior to discussion): IV1 indicated greater confidence in this MT than for Example 1, seemingly due to its more cohesive reporting structure and the perceived accuracy of the terminology and/or lexis used: *'[...] it makes sense, what I'm reading in English'* and *'the words that are translated seem more professional [...] it seems a proper report, to be honest with you [...] I would believe in this'*. On independently focusing on 'DRS', they noted that they neither understood the abbreviation nor were able to decipher its meaning by reversing the order of the letters. Despite this, the MT was assigned *'a confident four'*.

Discussion: While examining the ST, IV1 initially misread the patient as having been referred to Psychiatry on the basis of a cardiac cause having been ruled out, rather than its having already been determined that they had no known psychiatric history. On clarifying this point during the discussion stage, IV1 speculated on the possible meaning of 'DRS': *'I don't know exactly what it stands for, but I would say something on those lines of anxiety, panic attacks, that are still showing the pain and the pressure on the chest, giving the idea of a heart attack or a stroke, for example'*. However, as no such psychiatric history was present, IV1 ultimately remained unsure as to its significance.

Q3 (confidence scoring following discussion): IV1 indicated that they would not feel confident using the MT owing to the unresolved 'DRS': *'[...] in health, there's no assumptions, you need to have facts [...] it needs to be clear'*. They further added that *'[...] I do not trust Google Translate, because you are quite confident in most of the text and then it comes [to] the bottom and you think, 'Oh...' [...] 'Can I trust the rest of what is above? Maybe not'*. Notably, due to this ambiguity surrounding the patient's ongoing symptoms, IV1 not only revoked their previous score but *'would go even lower, because in this situation, it could be [a] more serious outcome [...] [she] still has the pain, so that needs to be investigated. And that is not clear'*. Their score was reduced from four ('confident') to two ('somewhat confident') on the basis that they no longer trusted the MT: *'[...] that took my confidence completely away [...] it makes sense what is above, but does it? [...] I don't trust it anymore'*.

Participant 2 (IV2): Advanced Clinical Practitioner

IV2, a native English speaker, had worked extensively in emergency care nursing in the NHS for nearly thirty years across a variety of specialties and care settings and had since begun practising as an Advanced Clinical Practitioner¹. In terms of language knowledge, they indicated being able to ‘*make out bits and bobs that [...] you could pick up on*’ in Spanish but were otherwise unable to speak or read it. Throughout the interview, they appeared to recognise French words with greater ease, though did not claim to have any formal knowledge of the language. Both texts were presented to IV2 in the context of a patient in need of treatment, accompanied by medical record information from their home country.

Example 1: ‘AO’

Q2: IV2 immediately identified ‘AO’ as unrecognised but attributed this to unfamiliarity with the specialty in question rather than a translation issue (*‘I don’t know what ‘AO’ is [...] because I’m not an Ophthalmology specialist’*), noting that they would ‘*refer to a specialist to be on the safe side*’.

Q3 (confidence scoring prior to discussion): IV2 assigned the MT a score of two (‘somewhat confident’) on the basis that it enabled them to establish a rough clinical picture, though this was hampered by their own unfamiliarity with the specialty in question. Interestingly from a comparative confidence perspective and in view of the subsequent cardiological French ‘DRS’ example, they also noted that *‘If it was something like a cardiac presentation, for example, I’d know an awful lot more, and I could be a lot more confident in interpreting the text based on what I’d seen with the patient’*.

Discussion: IV2 indicated that their unfamiliarity with the specialty drew their attention to ‘AO’ and informed their decision to seek help with the MT, though they still perceived this ambiguity as being clinical in nature, rather than a translation issue: *‘[...] I thought ‘I don’t know what that means’, but Ophthalmology would look at that and [...] probably make an educated guess [...]’*. Furthermore, on discussing the lack of preceding full term with which to decipher ‘AO’ by comparison with other abbreviations in the text, IV2 highlighted the fact that *‘If there’s no context to it [...] you just don’t know what it is’*.

Q3 (confidence scoring following discussion): IV2 indicated that they would retain their confidence score of two (‘somewhat confident’), seemingly on the basis of most of the MT enabling them to establish a broad clinical picture. However, they emphasised having taken note of the ambiguity arising from the lack of context preceding ‘AO’ in their assessment: *‘[la presión intraocular] you can, sort of, interpret that, and see [...] that relates to the words before, but just the abbreviation, that has no context whatsoever, does it?’*.

¹ Health Education England (HEE) defines Advanced Clinical Practice as being ‘delivered by experienced, registered health and care practitioners. It is a level of practice characterised by a high degree of autonomy and complex decision making. This is underpinned by a master’s level award or equivalent that encompasses the four pillars of clinical practice, leadership and management, education and research, with demonstration of core capabilities and area specific clinical competence’. [10]

Q4 (potential risk to patient safety scoring following discussion): The MT was assigned a score of two ('capacity to contribute to or result in temporary harm requiring intervention') based on the need to seek assistance with the translation to avoid negatively impacting on patient safety, adding that *'you would definitely need to get help with this'*.

Example 2: 'DRS'

Q2: IV2 immediately identified 'DRS' as unrecognised, along with the inconsistent translation of gender throughout the MT (alternating between male and female) and cardiology-specific terminology with which they were unfamiliar ('resténose'/'restenosis'). Again, they indicated that the MT may be useful for obtaining valuable information but would be mindful of the errors identified in assessing its reliability for clinical purposes: *'[...] it does give me important information, but there are certain things in there that I would think, 'That's not right''*. They emphasised the importance of integrating the information provided by the MT into a wider clinical assessment of the patient and any attendant investigations, approaching the text with a degree of caution accordingly: *'[...] I would take some of the information from this, but I would have definitely got it interpreted and maybe referred to Cardiology if I was concerned about the patient, but [...] I wouldn't rely on it too heavily, at all'*.

Q3 (confidence scoring prior to discussion): Despite the untranslated 'DRS', a confidence score of two ('somewhat confident') was assigned to the MT on the basis of its enabling IV2 to establish an overall picture of the patient's medical history, though its use in practice would fit within the wider framework of their clinical assessment of the patient.

Discussion: IV2 indicated that despite obtaining a general overview of the clinical context from the MT, the non-translation of 'DRS' remained centrally problematic to treating the patient's symptoms and would therefore need to be resolved due to its potential criticality in clinical terms: *'I would want help with that, because I don't know what it is, and it might be very important [...] ['DRS' has] been investigated, nothing found, but what is it?'*

Q3 (confidence scoring following discussion): IV2 again expressed an appreciation for the clinical background provided by the MT, though within the context of its being complementary to their wider clinical assessment: *'[...] you would have confidence, in a way, that you're dealing with a certain situation'*. Despite this, their score was reduced from two ('somewhat confident') to one ('slightly confident') in acknowledgement of the need to resolve 'DRS', for which they indicated referring to a French-speaking colleague or Language Line [an interpreting and translation service used across the NHS].

Q4 (potential risk to patient safety scoring following discussion): IV2 assigned the MT a score of three ('capacity to contribute to or result in death') owing to the potentially significant nature of the medical issue at hand and the ambiguity surrounding the centrality of 'DRS' to determining the ongoing symptoms and prioritising patient safety: *'Because it's a cardiac presentation [...] [it's] probably more critical to understand the interpretation because it could be something that you could get very wrong'*.

In closing, IV2 indicated that the interview had given them pause in considering the use of abbreviations in their own professional practice, particularly given the ubiquity of such features in both verbal and written healthcare contexts. However, time was noted as a contributor to their continued use: *'We use abbreviations all the time at work [...] and if I was to document, should I be using abbreviations or should I be using the full text? And I think I should be using the full text. But time is a factor [...] [Abbreviation use] is a language in itself'*.

7 Discussion of Preliminary Findings

7.1 Impact of MT Use with Patient Medical Record Documentation

The examples analysed in Section 4 indicate that certain medical abbreviations, particularly those for which no surrounding context in the form of a preceding full term is provided in the ST, run the risk of remaining entirely untranslated in the corresponding MT output. In the case of 'AO', a correct translation should have reflected the directly equivalent abbreviation available in either Latin or English. By contrast, for 'DRS', the failure of the MT to provide an appropriate translation may be attributed to the fact that no such directly equivalent abbreviation exists in English, in addition to the absence of preceding context in the ST. Given the higher-stakes setting and the centrality of both abbreviations to comprehension of their respective clinical contexts, their non-translation not only implies an error at the linguistic level but also a potential increase in risk to patient safety.

7.2 Awareness among Healthcare Staff of Risks Posed by MT Use with Patient Medical Record Documentation

While both participants initially underestimated the relevance of the selected abbreviations to the MT's clinical fitness for purpose, their respective responses indicated an interesting difference in reasoning as to the influences on their perceptions in this regard. IV1's view of MT reliability appeared to be more heavily influenced by the perceived text quality at a broader linguistic level and its conformance to typical reporting structures, as well as the cohesiveness of each text as a whole. Further to this, they acknowledged having originally overlooked the untranslated 'AO' owing to the assumption that it represented a different type of information. IV2's view of this reliability appeared to depend more heavily on their familiarity with the specific specialty, rather than an awareness of a translation issue present in the MT. This was reflected in the fact that, while IV2 immediately identified both abbreviations as unrecognised, their initial assumption with regard to 'AO' was that this was due to their unfamiliarity with ophthalmology rather than its remaining untranslated in the MT output. Following discussion, both participants appeared to adopt a more critical eye from a translation perspective, with IV1 in particular paying closer attention to other untranslated abbreviations in Example 1 and both IV1 and IV2 focusing more closely on the impact of the lack of a preceding full term on the (non-)translation of both 'AO' and 'DRS'.

In terms of changes to participants' degree of risk awareness following discussion, IV1 initially expressed less concern over the non-translation of 'AO'. This was seemingly based on their perception of the patient's overall clinical stability. However, following discussion of 'DRS', they indicated a significant drop in trust in the MT output owing to the potentially

serious clinical implications arising from its non-translation, this distrust extending to the reliability of the entire MT as a result. By comparison, IV2 appeared to adopt a more integrative approach to assessing risk in the MT output in terms of its allowing them to establish an overall picture of the patient's background within the framework of their wider clinical assessment. However, at no point were the errors and/or ambiguities present in the MT dismissed by IV2 as a result of this approach; to the contrary, it was emphasised throughout that these would need to be identified and addressed in order to preserve patient safety, particularly given the ambiguous nature of the ongoing symptoms indicated by 'DRS'.

In regard to management of the potential risks posed by using the MT in a professional setting, both participants demonstrated sound awareness of the need to confirm the content of the MT to ensure clarity of comprehension and preserve patient safety, whether through deferral to a clinical specialist and/or professional language services. In particular, IV2 emphasised the fact that while they considered the MT to contain potentially important information that should not be disregarded outright based on the ambiguities and/or errors present, the MT output would require both double-checking and integration into their own clinical assessment of the patient, rather than relying on it in isolation. Furthermore, IV2 commented that the interview had inspired them to reconsider their own practices in terms of abbreviation usage in a real-life setting, despite the time constraints present in their work environment.

8 Conclusions and Future Work

Both participants demonstrated an encouraging degree of risk awareness in their respective approaches to managing ambiguities and/or errors present in patient medical record information translated with MT. Nonetheless, it would appear that there remains a need to raise awareness among healthcare staff concerning the closer consideration of unrecognised medical abbreviations in MT output, particularly where these are inadequately translated or not translated at all, and may increase the potential for risk to patient safety as a result. This is especially relevant to medical abbreviations lacking in preceding context in the ST, the non-translation of which may impact negatively on care delivery due to their being overlooked by healthcare staff as less immediately relevant to clinical assessment than more readily recognisable medical terminology, misinterpreted as other types of information, or simply assumed to belong to another specialty. The difference in reasoning as to the influences on each participant's perceptions of the MT's fitness for purpose also raises interesting questions as to the degree to which future participants may interpret text quality at the broader linguistic level (that is, in terms of lexis and/or terminology, conformance to expected genre conventions, etc.), along with the extent to which this perception influences their awareness of the risks presented by both medical abbreviations and their (non-) translation with MT. These findings are currently being followed up by interviews with a further 18 participants and shall be subject to more comprehensive analyses accordingly.

9 References

1. Albrecht, U-V., Behrends, M., Schmeer, R., Matthies, H. K., von Jan, U.: Usage of Multilingual Mobile Translation Applications in Clinical Settings. *JMIR mHealth and uHealth*, 1(1), e4 (2013).
2. BBC News, <http://news.bbc.co.uk/1/hi/health/7171453.stm>, last accessed 30/04/23.
3. Bell, S., Saliba, V., Evans, G., Flanagan, S., Ghebrehewet, S., McAuslane, H., Sibal, B., Mounier-Jack, S.: Responding to measles outbreaks in underserved Roma and Romanian populations in England: the critical role of community understanding and engagement. *Epidemiology and Infection*, 148, 1-8 (2020).
4. Bouillon, P., Gerlach, J., Spechbach, H., Tsourakis, N., Halimi, M., Ismahene, S. BabelDr vs Google Translate: a user study at Geneva University Hospitals (HUG). In: 20th Annual Conference of the European Association for Machine Translation (EAMT) (2017).
5. Brunetti, L., Santell, J. P. and Hicks, R. W.: The Impact of Abbreviations on Patient Safety. *The Joint Commission Journal on Quality and Patient Safety*, 33(9), 576–583 (2007).
6. Chaudary, R. and Herrera, H.: Language barriers and patient care in community pharmacy. *International Journal of Pharmacy Practice*, 24(Supplement 3), 106–115 (2016).
7. Chen, X., Acosta, S. and Barry, A. E.: Evaluating the Accuracy of Google Translate for Diabetes Education Material. *JMIR Diabetes*, 1(1), 1–11 (2016).
8. Davis, S. H., Rosenberg J., Nguyen J., Jimenez M., Lion K. C., Jenicek G., Dallmann H., Yun K.: Translating Discharge Instructions for Limited English–Proficient Families: Strategies and Barriers. *Hospital Pediatrics*, 9(10), 779–787 (2019).
9. Grabar, N., Dalloux, C. and Claveau, V.: CAS: corpus of clinical cases in French. *Journal of Biomedical Semantics*, 11, 7 (2020).
10. Health Education England (HEE), <http://www.hee.nhs.uk/our-work/advanced-clinical-practice/what-advanced-clinical-practice>, last accessed 18/06/23.
11. Hwang, K., Williams, S., Zucchi, E., Chong, T. W. H., Mascitti-Meuter, M., LoGiudice, D., Goh, A. M. Y., Panayiotou, A., Batchelor, F.: Testing the use of translation apps to overcome everyday healthcare communication in Australian aged-care hospital wards - An exploratory study. *Nursing Open*, 9(1), 578-585 (2022).
12. Institute for Safe Medication Practices, <https://www.ismp.org/resources/please-dont-sleep-through-wake-call>, last accessed 30/04/23.
13. Kilgarriff, A., Baisa, V., Bušta, J., Jakubiček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V.: The Sketch Engine: ten years on. *Lexicography*, 1, 7-36 (2014).
14. Leite, F. O., Cochat, C., Salgado, H., da Costa, M. P., Queirós, M., Campos, O., Carvalho, P.: Using Google Translate[®] in the hospital: A case report, *Technology and Health Care*, 24(6), 965–968 (2016).
15. Miranda-Escalada, A., Gonzalez-Agirre, A. and Krallinger, M.: Overview of automatic clinical coding: annotations, guidelines, and solutions for non-English clinical cases at CodiEsp track of CLEF eHealth 2020. In: CEUR Workshop Proceedings, 2696, 22–25 (2020).
16. Moberly, T.: Doctors choose Google Translate to communicate with patients because of easy access. *British Medical Journal*, 362 (2018).
17. NHS England, <https://www.england.nhs.uk/wp-content/uploads/2018/09/guidance-for-commissioners-interpreting-and-translation-services-in-primary-care.pdf>, last accessed 30/04/23.
18. Noack, E., Kleinert, E., and Müller, F.: Overcoming language barriers in paramedic care: a study protocol of the interventional trial ‘DICTUM rescue’ evaluating an app designed to improve communication between paramedics and foreign-language patients. *BMC Health Services Research*, 20(1), 223 (2020).

19. Otero López, M. J., Martín Muñoz, R. and Domínguez-Gil Hurlé, A.: SEGURIDAD DE MEDICAMENTOS: Abreviaturas, símbolos y expresiones de dosis asociados a errores de medicación. *Farmacia Hospitalaria*, 28(2), 141–144 (2004).
20. Panayiotou, A., Gardner, A., Williams, S., Zucchi, E., Mascitti-Meuter, M., Goh, A. M. Y., You, E., Chong, T. W. H., LoGiudice, D., Lin, X., Haralambous, B., Batchelor, F.: Language Translation Apps in Health Care Settings: Expert Opinion. *JMIR mHealth and uHealth*, 7(4), e11316 (2019).
21. Panayiotou, A., Hwang, K., Williams, S., Chong T. W. H., LoGiudice, D., Haralambous, B., Lin, X., Zucchi, E., Mascitti-Meuter, M., Goh, A. M. Y., You, E., Batchelor, F.: The perceptions of translation apps for everyday health care in healthcare workers and older people: A multi-method study. *Journal of Clinical Nursing*, 29(17-18), 3516-3526 (2020).
22. Patil, S. and Davies, P.: Use of Google Translate in medical communication: evaluation of accuracy. *British Medical Journal*, 349 (2014).
23. Royal College of Midwives: Stepping up to Public Health: A new maternity model for women and families, midwives and maternity support workers. Royal College of Midwives, London (2017).
24. Sheik-Ali, S., Dowlut, N., McConaghie, G.: Breaking down language barriers with technology. *The Bulletin of the Royal College of Surgeons of England*, 98(3), 138-140 (2016).
25. Sheppard, J. E., Weidner, L. C. E., Zakai, S., Fountain-Polley, S., Williams, J.: Ambiguous abbreviations: an audit of abbreviations in paediatric note keeping. *Archives of Disease in Childhood*, 93(3), 204–206 (2008).
26. Starlander, M., Bouillon, P., Rayner, M., Chatzichrisafis, N., Hockey, B. A., Isahara, H., Kanzaki, K., Nakao, Y., Santaholma, M.: Breaking the language barrier: machine assisted diagnosis using the medical speech translator. *Studies in health technology and informatics*, 116, 811-816 (2005).
27. Taira, B. R., Kreger, V., Orue, A. and Diamond, L. C.: A Pragmatic Assessment of Google Translate for Emergency Department Instructions. *Journal of General Internal Medicine*, 36(11), 3361–3365 (2021).
28. The Joint Commission, <https://www.jointcommission.org/resources/news-and-multimedia/fact-sheets/facts-about-do-not-use-list>, last accessed 30/04/23.
29. Turner, A. M., Choi, Y. K., Dew, K., Tsai, M-T., Bosold, A. L., Wu, S., Smith, D., Meischke, H.: Evaluating the Usefulness of Translation Technologies for Emergency Response Communication: A Scenario-Based Study. *JMIR Public Health and Surveillance*, 5(1), e11171 (2019).
30. Vieira, L. N., O’Hagan, M. and O’Sullivan, C.: Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11), 1515–1532 (2020).
31. Vieira, L. N., O’Sullivan, C., Zhang, X. and O’Hagan, M.: Machine translation in society: insights from UK users. *Language Resources and Evaluation* (2022).

10 Appendices

Interview questions

Q1 (ST only): What kind of information in the documentation, if any, would you be looking to identify, and for what purposes?

Q2 (MT with ST available for reference): If you identified, or thought you may have identified, an error in the machine translation output, what would you do?

Q3 (confidence scoring requested from all participants prior to and following discussion): How confident would you feel about using this machine translation output in a real-life professional situation? Please give a score for your degree of confidence and explain your answer.

Q4 (potential risk to patient safety scoring requested from medically qualified participants only, following discussion): Please provide a score for the degree of potential risk to patient safety you would consider this machine translation output to pose if used in a real-life professional situation. Please explain your answer.

Translations and Open Science

Exploring how translation technologies can support multilingualism in scholarly communication

S. Fiorini¹, A. Tezcan², T. Vanallemeersch³, S. Szoc³, K. Migdisi³, L. Meeus³,
and L. Macken²

¹ OPERAS, Belgium

² Ghent University, Belgium

³ CrossLang, Belgium

susanna.fiorini@operas-eu.org

Abstract. English is by and large considered as the *lingua franca* of scholarly communication. Such a generalised use has certainly the advantage of facilitating international exchanges, but it also generates inequalities among researchers, and limits the dissemination of scientific knowledge. Although translation could be promptly identified as the solution, scholarly communication has historically been marked by a shortage of human and financial resources to support traditional translation processes. The goal of this paper is to present a multi-user approach to machine translation evaluation for a use in scholarly communication. In particular, the paper introduces the fine-tuning and evaluation methodology set up to comply with the needs of different target user *personas* (translators, researchers, readers). Given the focus of the conference, the paper will describe in more detail the evaluation methodology related to the “Translator” *persona*. The paper will also include general preliminary conclusions, and information about the on-going evaluation work.

Keywords: translation technology · machine translation · machine translation evaluation · multilingualism · open science.

1 Introduction

English is by and large considered as the *lingua franca* of scholarly communication. Such a generalised use has certainly the advantage of facilitating exchanges in an increasingly internationalised research landscape. However, this linguistic dominance also generates inequalities among researchers [1], and limits the dissemination of scientific knowledge within non-English speaking communities [2, 3]. In this context, translation could be promptly identified as a solution to help eliminate language barriers and inequalities in research, according to open science principles. Yet, scholarly communication has historically been marked by a shortage of human and financial resources to support traditional translation processes.

The *Translations and Open Science* project was launched to promote a more structured implementation of translation technologies [4, 5] in order to foster

language equality in scholarly communication as well as the dissemination of knowledge at lower costs and with greater efficiency. The expected deliverable of the project is a technology-based scientific translation service, combining technology tools, digital language resources and human skills. Besides in-domain multilingual data and collaborative translation features, the service is intended to provide scientific translators and researchers with adapted machine translation engines, which will serve different purposes and usage scenarios. A use-case study [6] conducted as part of the project suggested indeed that machine translation is used not only by translators as a productivity aid, but also by researchers as a foreign-language writing assistant, as well as by readers of different profiles, who leverage machine translation for discoverability and gisting purposes.

The goal of this paper is to present our approach to machine translation evaluation for use in scholarly communication. In particular, we will introduce the methodology we set up in order to fine-tune and evaluate machine translation engines, while taking into account the different target user profiles and the associated needs. Given the focus of the conference, the paper will describe in more detail the evaluation methodology related to the “Translator” *persona* and usage scenario. The paper will also include general preliminary conclusions, as well as information about the on-going evaluation work.

2 Machine translation for scholarly communication

The conducted use-case study [6] allowed us to draft an overview of the current translation practices in scholarly communication across a variety of scientific domains. Based on a series of interviews and workshops involving a total of 30 participants⁴, the study revealed different levels of acceptance of translation technologies, and in particular machine translation, among scientific translators and researchers according to their domain of specialisation. Although these differences can be partially explained by the very specific characteristics of disciplinary content and writing standards, we also observed an impact of the subjective user attitudes on the acceptance of machine translation. To rely on objective data, not affected by the enthusiasm or the scepticism of users, we decided to carry out an *ad hoc* evaluation in order to assess the relevance of machine translation deployment in scholarly communication.

Dataset collection for machine translation fine-tuning The use-case study showed that more than 80% of the interviewed translators and researchers who use machine translation work with free, generic online engines. For our evaluation, we decided to assess whether fine-tuning can help to produce better machine translation output, especially by taking into account specialised terminology, which is crucial in scientific texts. In order to do so, we collected in-domain parallel language datasets in the English-French language pair in three

⁴ 15 scientific translators, 5 researchers, 8 academic publishers, 1 academic librarian, 1 translation technology engineer, all scientific domains combined.

pilot scientific domains. The pilot domains and the characteristics of the collected resources are the following:

1. Climatology and Climate Change (Physical Sciences): 100,563 collected segments, 397 extracted⁵ terms; translation direction of the bilingual corpus and evaluation task → English to French;
2. Neurosciences (Life Sciences): 103,175 collected segments, 415 extracted terms; translation direction of the bilingual corpus and evaluation task → English to French;
3. Human Mobility, Environment, and Space (Social Sciences and Humanities): 112,963 collected segments, 300 extracted terms; translation direction of the bilingual corpus and evaluation task → French to English.

As an example, the Human Mobility, Environment, and Space corpus was split into four subsets: a training set (104,539 segments), a validation set (1,896 segments), a test set for automatic evaluation (2,183 segments), and an evaluation set for human evaluation (4,345 segments, including 954 segments specially selected and collected for human evaluation). A similar approach was used for the two other domains in order to proceed to the evaluation tasks.

Choice of engines to be evaluated Since openness is a core principle of the *Translations and Open Science* project, we primarily considered open-source engines, allowing for customisation. We decided to pick two engines presenting different customisation methods: an engine based on an open-source library with highly customisable, multi-parameter setup (OpenNMT), and an engine allowing for simplified, user-level adaptation (ModernMT). In this way, we wanted to be able to determine what kind of fine-tuning effort (if any) is necessary in order to produce better output with disciplinary texts.

Although it does not comply with the open-source requirement, we also included in the evaluation the engine which is, according to our use-case study, the most used by our target community (DeepL).

Fine-tuning To train and fine-tune the engine based on the OpenNMT library, we started with a from-scratch training on open-source parallel datasets provided in OPUS [7]. This resulted in a generic machine translation model, which we then fine-tuned on the collected specialised datasets.

Concerning ModernMT, we fine-tuned the baseline engine by uploading the collected corpora in TMX format through the dedicated feature provided in the online user interface.

With regard to DeepL, we fine-tuned the baseline engine through the *Glossary* feature for terminology customisation. However, it should be noted that, according to our use-case study, most of our target users do not work with this feature, which is also not supported yet in all the API configurations available.

⁵ Automated term extraction from the collected corpora, with human review.

These points could be potential limitations in the case of a large-scale deployment, so they will have to be taken into account together with the evaluation outcomes.

3 Evaluation of machine translation for scholarly communication

After the fine-tuning, we proceeded to an in-domain evaluation of the selected machine translation engines. According to the use cases identified in our previous research, the evaluation was set up to provide information about the usability of the raw machine translation output generated by the evaluated engines in the three scenarios below:

1. A researcher using machine translation as a support to write a paper in a foreign language or to translate a paper into a foreign language;
2. A translator using machine translation to perform post-editing in a computer-assisted translation environment;
3. A reader using machine translation to get an idea of the content of a scientific publication.

The scores relating to specialised terminology compliance in machine translation output are also leveraged to understand whether raw machine translation can be useful to automatically translate publication metadata and therefore improve the discoverability of research in multiple languages.

Automatic evaluation The engines were submitted for automatic evaluation by producing output for in-domain test datasets with both baseline and fine-tuned engines. The outputs produced by the six engines (three baseline and three fine-tuned engines) were compared to reference translations using automatic evaluation metrics such as the statistical metrics BLEU and TER (Translation Edit Rate) and the neural (deep learning based) metric COMET [9]. The MATEO software [10] was used to calculate these metrics. The comparison between the baseline and fine-tuned engines was intended to provide further insight into fine-tuning needs, and in particular to bring additional information about the relevance and the required level of fine-tuning effort in order to improve machine translation output.

Besides calculating metric scores, we also visualised the differences between machine translation outputs. The software used shows the difference on character level between the reference translation and the machine translation output, as well as the character-based edit distance between the two sentences.

Human evaluation As part of the human evaluation task, we evaluated the output of the three machine translation engines which obtained the best scores in automatic evaluation. The evaluation was set up to assess machine translation output usability for the three following *personas* and usage scenarios:

Persona 1 - “Translator”: professional translator who masters the source language, is a native speaker of the target language, and has a good knowledge of the domain in question. This *persona* performed an adequacy assessment task, as well as a post-editing task in a dedicated evaluation tool.

Persona 2 - “Expert”: researcher specialised in the domain in question, who uses machine translation to (a) translate their scientific publication, (b) write an article in the target language (writing aid), or (c) gist scientific texts that are not written in their native language (reading aid). Having a good to native knowledge of the source and target languages, as well as a perfect command of specialised terminology in both languages, this *persona* performed the same evaluation tasks assigned to the “Translator” *persona*: adequacy and post-editing (see section 3.1.3).

Persona 3 - “Layperson”: a person who has at most basic knowledge in the domain (e.g. a non-academic reader or a researcher in a different scientific domain). This *persona* has good to excellent knowledge of the target language and makes use of machine translation to gist educational scientific texts. The participants to this task read text excerpts of 100-200 words, drawn from the evaluation set, in a cumulative self-paced reading view. Based on text characteristics - such as the origin of the excerpt (abstract or full text), sentence length, and lexical variety - the texts were classified into different sets which were submitted to different user groups. The human reference translation was used as a benchmark. Reading time was measured. After reading each excerpt, the users were asked to answer multiple-choice comprehension questions as an incentive to read the text attentively.

The “Translator” evaluation setup Given the focus of the conference, we only present in detail the evaluation methodology for the “Translator” *persona* and usage scenario. As part of this human evaluation subtask, two professional translators, specialising in the domains in question, performed for each domain adequacy and post-editing tasks.

The adequacy task consisted in judging the adequacy of the machine translated segments (sentences) of scientific publications, by assigning a score between 1 and 5. The aim of this task was to assess how adequately the machine translation of the segment expressed the source segment’s meaning, and, by consequence, how useful the translation was for gisting and discoverability purposes.

Around 500 segments extracted from scientific papers, reviews and abstracts were shown to each evaluator in the order they appear in the document. For each segment, the evaluators were provided with: (1) the part of the paragraph preceding the evaluated segment, (2) the segment itself, (3) the remainder of the paragraph, and (4) machine translation outputs, randomly ordered to avoid evaluator bias (in this way, the evaluators did not only judge the overall quality of machine translation outputs, but also ranked them implicitly). The evaluators were also provided with the abstracts of the documents from which the evaluation segments originate in order to provide more context. Reference translations were not shown to avoid bias.

The post-editing task consisted in asking the evaluators to produce a publishable translation (a terminologically valid, grammatically correct, fluent translation conveying the meaning of the source sentence), based on a source segment, its context, and a machine translation output. The evaluator was also asked to provide a score for perceived post-editing effort for each segment. This task was performed on a different test set than the one used for the adequacy task. As in the adequacy task, around 500 segments were shown, without reference translation, and in the order they appear in the document. However, only one machine translation output for each segment was provided (the evaluators were provided with output from different engines without knowing which engine had been used to translate a specific source segment).

Three metrics were applied to assess the productivity with each machine engine:

1. temporal effort (average time per word);
2. technical effort based on human-target TER (HTER) scores via measurement of post-editing difference (PEDiff) between machine output and the translation produced by the evaluator;
3. perception of effort (see above).

Samples drawn from the post-edited outputs were annotated using the MQM framework. Error annotations were performed using the seven high-level error dimensions: terminology, accuracy, linguistic conventions, style, locale conventions, audience appropriateness, and design and markup. The output edited as part of the “Expert” *persona* setup were also annotated according to the same standards, in order to determine the relations between error types and editing behaviours based on user profiles (for instance, determining whether one *persona* is more likely to correct terminology errors rather than style).

4 Preliminary conclusions and work in progress

According to the use-case study conducted as part of the *Translations and Open Science* project, machine translation is frequently used to produce multilingual publications in some domains (100% of the life sciences and physical sciences researchers and translators interviewed use machine translation), while in other domains we observed a more reluctant attitude towards this technology (20% of the humanities and social sciences researchers and translators interviewed use machine translation). This data suggests that, besides the required technical efforts, investments in training and literacy programs are also needed in order to efficiently deploy translation technologies, and in particular machine translation, in scholarly communication.

The collection of bilingual scientific datasets for machine translation fine-tuning also raised various challenges. Firstly, the amount of bilingual data available in scientific publications is limited. As we said, translation is not a systematic activity in scholarly communication due to disciplinary standards and

a shortage of resources. Moreover, in most of the cases only abstracts are translated, the full text papers being only available in one language⁶. Also, it can be difficult to determine the origin of the translation, which means that it is not always possible to easily identify and exclude low quality-translations or *translationese* from the test sets. Secondly, a considerable portion of the bilingual data we found is published under licences which expressly forbid data collection and processing, or which do not provide clear information about the authorised uses (~40% of the identified data sources). We were mainly able to collect and process the publications under Creative Commons licences, according to the conditions established by the applied licence type. For the remaining data, in a few cases we received the authorisation to collect and process data from the right owners, otherwise we relied on the Text and Data Mining exception (TDM), introduced by the European directive 2019/790 and transposed into French law in 2021. When the publication did not fall under the TDM exception and we could not get the required authorisations, we simply did not collect any data (~30% of the identified data sources). Finally, we observed a general lack of standardisation among data sources when it comes to formats and keyword classification of scientific publications, which can complicate data collection through automated processes.

As for the evaluation task, the automatic evaluation results seem to show that there is no significant improvement in the performance of the fine-tuned versions of DeepL and ModernMT, while the OpenNMT engine does perform better after fine-tuning with the specialised datasets collected as part of the *Translations and Open Science* project, and performs even better after also adding the SciPar corpus [11], which contains parallel corpora from scientific abstracts, all domains combined (Fig. 1-3). However, the overall performance of the OpenNMT engine remains lower than DeepL and ModernMT even after fine-tuning, except for the “Thesis abstracts” document type in the SH7 discipline as well as for the “Review abstracts” and “Thesis abstracts” document types in the LS5 discipline.

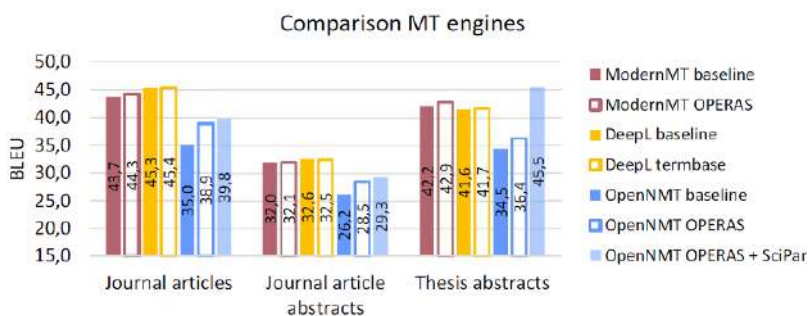


Fig. 1. BLEU scores by engine and document type for the SH7 discipline

⁶ Out of the 23 sources from which we collected data, only 9 had some full text papers translated.

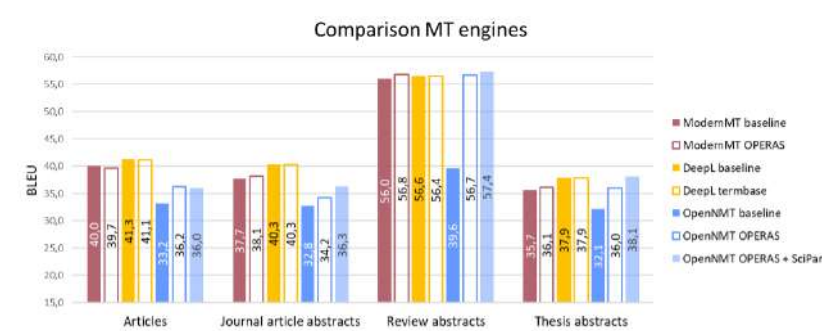


Fig. 2. BLEU scores by engine and document type for the LS5 discipline

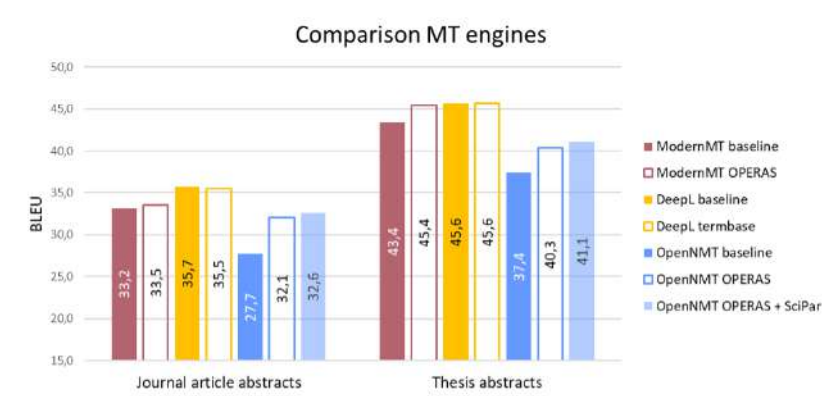


Fig. 3. BLEU scores by engine and document type for the PE3_10 discipline

At a first glance, these results seem to suggest that, given the effort required to collect and prepare parallel data, fine-tuning might not be the most effective strategy to improve machine translation output. We even found cases where the baseline DeepL engine performed better than the fine-tuned one: however, given that fine-tuning in DeepL only covers terminology, this could be due to terminology inconsistencies in the reference translations.

In the light of the evaluation outcome after adding the SciPar corpus to the OpenNMT engine, another hypothesis is that data collection for fine-tuning should not be strictly narrowed to in-domain texts only. This is key information for the general sustainability of our approach.

The human evaluation performed by professional translators overall confirmed the ranking established by the automatic evaluation: for the three disciplines, DeepL tends to have on average the lowest post-edit time and perceived effort as well as the highest user rating in adequacy tasks, followed by

ModernMT and OpenNMT. The MQM annotation results in the same machine translation engine ranking, at least for disciplines SH7 and LS5 (data for the PE3_10 discipline is still under production at the time the present paper is being drafted). Given the importance of specialised terminology in scientific texts, we focused our analysis on terminology errors, which might discourage the use of raw machine translation to automatically translate publication metadata for discoverability purposes (Fig. 4 and 5).

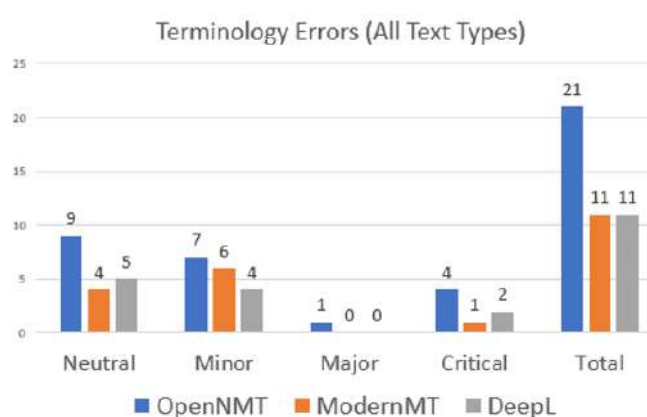


Fig. 4. Terminology errors for SH7 discipline

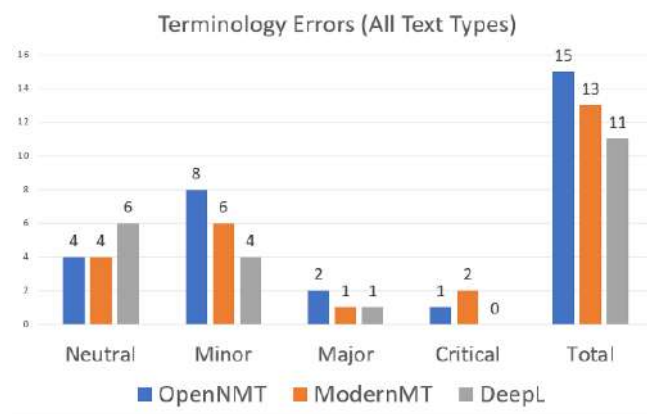


Fig. 5. Terminology errors for LS5 discipline

When it comes to correlations between translators regarding post-edit time, perceived effort, HTER and adequacy ratings (more specifically, Pearson product-

moment correlation coefficient⁷), they range from 35% to 45% in the SH7 discipline, from 35% to 52% in the LS5 discipline, and from 29% to 52% in the PE3_10 discipline. Regarding the machine translation engine ranking obtained, it is plausible to assume that DeepL may have benefited from the extensive use over time of its free version by researchers, while OpenNMT might have been more affected by the variable quality of fine-tuning data. In order to improve the reliability of the results in the future, the study could therefore benefit from the inclusion of a larger panel of evaluators per profile and a clearer view of the nature and quality of the data used for the fine-tuning of machine translation engines.

Acknowledgements The authors acknowledge F. Barbin and K. Hernandez-Morin (Université Rennes 2) for their contribution to the evaluation task of the *Translations and Open Science* project. The project is funded by the French National Fund for Open Science, which includes contributions from the the French Ministry of Higher Education and Research, universities and other research institutions. The project also received a special contribution from the French Ministry of Culture.

References

1. V. Ramírez-Castañeda, 2020, Disadvantages in preparing and publishing scientific papers caused by the dominance of the English language in science: The case of Colombian researchers in biological sciences. PLoS ONE 15(9): e0238372. <https://doi.org/10.1371/journal.pone.0238372>
2. Di Bitetti, Mario S., and Julián A. Ferreras, 2017, Publish (in English) or perish: The effect on citation rate of using languages other than English in scientific publications, *Ambio* 46.1: 121-127
3. Z. Taşkın, G. Doğan, E. Kulczycki, A. Zuccala, 2020, Science needs to inform the public. That can't be done solely in English, LSE blog. <https://blogs.lse.ac.uk/covid19/2020/06/18/long-read-science-needs-to-inform-the-public-that-cant-be-done-solely-in-english/>
4. L. Bowker, J. Ciro, 2019, Machine translation and global research: Towards improved machine translation literacy in the scholarly community, Bingley, UK: Emerald Publishing
5. S. Fiorini et al., 2020, Rapport du groupe de travail "Traductions et science ouverte", Comité pour la science ouverte. 44 p.
6. Use case study for a technology-based scientific translation service, report to be published
7. J. Tiedemann, 2012, Parallel Data, Tools and Interfaces in OPUS. Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), 2214-2218
8. Papineni et al., 2002, BLEU: a Method for Automatic Evaluation of Machine Translation Kishore Papineni, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318

⁷ A measure of linear correlation between two sets of data.

9. Rei et al., 2020, COMET: A Neural Framework for MT Evaluation Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)
10. B. Vanroy, A. Tezcan, L. Macken, 2023, MATEO: MACHine Translation Evaluation Online. In M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, ... H. 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).
11. D. Roussis et al., 2022, SciPar: A Collection of Parallel Corpora from Scientific Abstracts, Proceedings of the Thirteenth Language Resources and Evaluation Conference

Current evidence of post-editease: differences between post-edited neural machine translation output and human translation revealed through human evaluation

Michael Farrell¹[0000-0002-7138-6639]

¹ IULM University, Milan, Italy
michael.farrell@iulm.it

Abstract. The experiment reported in this paper is a follow-up to one conducted in 2017/2018. The new experiment aimed to establish if the previously observed lexical impoverishment in machine translation post-editing (MTPE) has become more marked as technology has developed or if it has attenuated. This was done by focusing on two n-grams, which had been previously identified as MT markers, i.e., n-grams that give rise to translation solutions that occur with a higher frequency in MTPE than is natural in HT. The new findings suggest that lexical impoverishment in the two short texts examined has indeed diminished with DeepL Translator.

The new experiment also considered possible syntactic differences, namely the number of text segments in the target text. However no significant difference was observed.

The participants were asked to complete a short questionnaire on how they went about their tasks. It emerged that it was helpful to consult the source language text while post-editing, and the original unedited raw output while self-revising, suggesting that monolingual MTPE of the two chosen texts would have been unwise.

Despite not being given specific guidelines, the productivity of the post-editors increased. If the ISO 18587:2017 recommendation of using as much of the MT output as possible had been strictly followed, the MTPE would have been easier to distinguish from HT. If this can be taken to be generally true, it suggests that it is neither necessary nor advisable to follow this recommendation when lexical diversity is crucial for making the translation more engaging.

Keywords: post-editease, machine translation post-editing, neural machine translation, human translation, human machine translation output evaluation, DeepL Translator, Google Translate.

1 Introduction

Several researchers have reported the existence of features of post-edited machine translation output (MTPE) that distinguish it from human translated text (HT), defined as post-editease. By way of example, Castilho et al. looked at literary texts Google-

translated from English into Brazilian Portuguese and found evidence for post-edited in one of the two texts examined [1]; Volkart et al. found that post-edited machine translation was not only lexically poorer than human translation, but also less dense and less varied in terms of translation solutions [12]; Toral found that MTPE was simpler and more normalized and had a higher degree of interference from the source language than HT [11]; and Castilho et al. found evidence of post-edited features, especially in light post-edited texts and in certain domains [2]. By contrast, on the other hand, Daems et al. found no proof of the existence of post-edited, either perceived or measurable [5].

The experiment described in this paper is a follow-up to an experiment carried out for two consecutive years (2017 and 2018) with postgraduate university students of translation (IULM University, Milan). In the previous experiment, half of the students did an unaided human translation (HT) and the other half post-edited machine translation output (MTPE). Comparison of the texts produced in 2017/18 showed that certain turns of phrase, expressions and choices of words occurred with greater frequency in MTPE than in HT (MT markers), making it theoretically possible to design tests to tell them apart. To verify this, the author successfully carried out one such test in 2018 on a small group of six professional translators [6].

The primary aim of the new experiment described in this paper was *not* to show that MTPE generally results in an increase in productivity, which is well documented elsewhere, but to see if it is still possible to detect MT markers in MTPE, despite the advances in MT technology since 2018, and if it is also still possible to distinguish MTPE from HT simply by comparing the number of these markers found in each kind of text. The students were also asked to provide various details of how they went about their tasks.

2 Design and methods

Two short extracts from English-language Wikipedia entries were taken for the experiment: one about Slovakia (262 words) and one about the Euromaidan civil unrest in Ukraine (263). Besides being the same genre as used in the previous experiment, Wikipedia articles were chosen since they are likely to be less challenging for an MT engine than classic works of literature but more problematic than the boilerplate-style texts which are often considered to lend themselves best to machine translation.

The first text was selected since it contained the bigram *there are* four times, and the other because it contained the monogram *people* (used as the plural of the word person and not as the singular noun meaning populace) six times. These were the first two short extracts that contained at least four examples of the chosen n-gram in the space of approximately 250 words that the author came across while searching randomly through Wikipedia.

These two n-grams had been identified in the 2017/18 experiment as among the best MT markers, that is n-grams which were translated with a highly statistically significantly greater number of correct translation solutions in HT than in MTPE, and *there*

are was the specific bigram used in the above mentioned successful test to distinguish HT from MTPE carried out in 2018 on a small group of six professional translators [6].

Forty-two postgraduate students of translation (IULM University, Milan) were divided into two groups and worked from English into Italian. Group A (21 students) translated the Slovakia text and post-edited the machine-translated Ukraine text, and group B (21 students) translated the Ukraine text and post-edited the machine-translated Slovakia text.

The pay-for version of DeepL Translator was chosen as the MT engine for this experiment for two main reasons:

1. The week before, the students had machine translated several short extracts (250 words approx.) from Wikipedia entries using different free online MT engines to compare the quality of their raw output, and an overwhelming majority had judged DeepL Translator to be the best for this genre (87%).
2. In a recent survey among professional translators [7], the MT engine most used by the respondents who declared that they use MT at some point in their workflow turned out to be DeepL Translator (183 users). Its nearest rival Google Translate was only chosen by just over half that number (93 users). The majority of DeepL Translator users surveyed (102) stated that they use the pay-for version.

The students were deliberately not given any post-editing guidelines but were told that they should transform the machine translated output into a text of the same quality as a human translation for publication (full post-editing). Both the post-editors and the translators were told that the task was urgent and should be completed in the shortest possible time without compromising on quality. They were also told that the objective of the experiment was to compare the average time taken for each task. They were not told beforehand that their final texts would be analysed for traces of post-edited text. The latter was in reality the primary reason for the experiment.

The students were allowed to use any dictionaries and reference material they liked, including Wikipedia itself, and even to ask for advice on individual problems from friends and colleagues not involved themselves in the experiment in a way that would not disturb the others, for example via WhatsApp. The intention was to recreate something as near as possible to normal working conditions. They were however instructed not to use MT engines in any way to prevent the translators from turning their task into a second post-editing assignment. This unfortunately goes against the aim of recreating real working conditions since it was found in the aforementioned recent survey that just over 69% of professional translators use MT in some way during their workflow, but not necessarily to translate the whole text for subsequent post-editing [7].

The files for translation and post-editing were provided as word processor documents and the translators and post-editors worked in Microsoft Word. The task was presented in this way so that the students were not influenced by the segmentation imposed by CAT or post-editing tools. It has in fact been observed that machine translation output normally has the same number of segments as the original language text, whereas translators who are working without a CAT tool sometimes organize the translated text into a different number of sentences. This can be verified by taking the first 26 sentences of Chapter 3 of *The Adventures of Pinocchio* by Carlo Collodi [3],

machine-translating them with Google Translate and comparing the output with the 1926 translation by Carol Della Chiesa [4]. The raw MT output is also organized into 26 sentences, whereas Della Chiesa’s translation has 28. It could be argued that it is rather obvious that there will be the same number of segments in a machine-translated text as in the original, but some machine translation engines today work at larger-than-segment level, notably ModernMT and possibly also DeepL Translator [7]. To make comparisons in this experiment, it was decided to count the number of segments created using the default segmentation rules of the two most used CAT tools according to the previously mentioned recent survey, Trados Studio and memoQ [7].

The students were also asked to complete a short questionnaire after they delivered their files to report some details of how they went about their tasks.

Unfortunately, one student misunderstood the instructions and translated and post-edited the same text; his work was discarded since the results of one of the two tasks were probably influenced by having done the other. Another student was not a native Italian speaker; her work was discarded since her translation choices may have been affected by her native language. Yet another student delivered a damaged file; it was however possible to evaluate the undamaged one. And one student did not deliver their files at all. In the end, 20 post-edited Ukraine texts, the same number of post-edited Slovakia texts, 19 translated Ukraine texts and 18 translated Slovakia texts were analysed.

Most of the variables measured in this paper are non-numeric, non-parametric, categorical variables which can only take on a limited number of values. For this reason, when possible, the widely used chi-square (χ^2) test was chosen for the statistical analyses. The significance level was set to .05, as per convention, to ensure a 95% confidence level, and the online chi-square test calculator provided by Jeremy Stangroom was used [9]. The results are reported in the format required by the American Psychological Association (APA) [10].

3 Results and discussion

3.1 Time comparison

As expected, and as is commonly reported, it took less time on average to post-edit the MT output than to translate the same text from scratch (Table 1).

Table 1. Translation and post-editing times in minutes and seconds.

Text	Length of text (words)	Translation time (mean \pm SD)	Post-editing time (mean \pm SD)	Productivity increase (%)
Slovakia	262	37:34 \pm 5:28	22:59 \pm 8:06	39.50%
Ukraine	263	38:47 \pm 10:00	20:01 \pm 7:18	47.99%

The productivity increase was calculated as the translation time minus the post-editing time, divided by the translation time. This was then multiplied by one hundred to obtain a percentage.

3.2 Slovakia text n-gram

The source text contained the bigram *there are* four times. DeepL Translator translated the bigram with *ci sono* three times and *vi sono*¹ once. Table 2 shows the translation solutions the translators and post-editors chose. The number shown is the overall number of occurrences of the n-gram indicated in all the texts of the given type (18 translated texts, 1 raw output and 20 post-edited texts). Since the number of texts in each category is different, the overall percentage number of occurrences should be considered when making comparisons.

Table 2. Italian translation solutions in HT, raw MT output and MTPE for *there are*

	Translation		Raw output		MTPE	
abbonda	1	1.39%				
ci sono	36	50.00%	3	75.00%	52	65.00%
è caratterizzata da	1	1.39%				
è possibile ammirare	1	1.39%				
è possibile trovare	1	1.39%				
è ricca di					1	1.25%
esistono	4	5.56%				
presenta	1	1.39%			1	1.25%
si possono trovare	7	9.72%				
si ritrovano					1	1.25%
si trovano	6	8.33%			1	1.25%
sono presenti	6	8.33%			5	6.25%
troviamo	1	1.39%				
vengono offerti	1	1.39%				
vi sono	5	6.94%	1	25.00%	19	23.75%
vi trovano	1	1.39%				
Totals	72	100%	4	100%	80	100%

From Table 2, it is evident that there is less variety in the solutions the post-editors chose since they are clearly primed by the raw output. This difference is statistically significant, as can be verified using the contingency table below (Table 3).

¹ Equivalent to *ci sono* but higher in register.

Table 3. Lexical variety contingency table for *there are*.

	Translation	MTPE
ci sono	36	52
vi sono	5	19
Other n-gram	31	9
$(\chi^2 (2, N = 152) = 22.82, p < .05)$		

The fact the raw output already contained two different translation solutions was unexpected (Table 2). Indeed, the same text translated by Google Translate contained the same solution all four times (*ci sono*). The presence of alternative translation solutions in DeepL Translator raw output is discussed in more detail below under *Degree of naturally occurring lexical variety in DeepL Translator raw output*.

In one of the previous experiments reported in 2018, a 273-word text containing five occurrences of *there are* was given to three professional translators for translation, and Google-translated and given to another three for full post-editing. None of the translators translated *there are* with *ci sono*, whereas all the post-editors left at least one occurrence of *ci sono*. Therefore, if the 2018 texts are split into two sets on the basis of how many times *ci sono* was chosen as the translation solution, it is possible to identify the MTPE with 100% accuracy. The same method of splitting the 2023 texts into two sets according to the number of occurrences of *ci sono* results in five misattributed texts. In other words, the translations are identifiable with $13/18 = 72.22\%$ accuracy and the MTPE, with $15/20 = 75\%$ accuracy.

3.3 Ukraine text n-gram

The source text contained the monogram *people*, used as the plural of the word *person*, six times. The raw output from DeepL Translator contained the monogram *persone* seven times since a demonstrative pronoun plus adjective in one of the source text sentences (*those killed*) was resolved into a noun plus adjective (*persone uccise*). Seven of the translators chose to do the same (Table 4). The number shown in Table 4 is the overall number of occurrences of the n-gram indicated in all the texts of the given type (19 translated texts, 1 raw output and 20 post-edited texts). Since the number of texts in each category is different, the overall percentage number of occurrences should be considered when making comparisons.

Table 4. Italian translation solutions in HT, raw MT output and MTPE for *people*

	Translation		Raw output		MTPE	
17	1	0.75%				
cittadini	1	0.75%				
coloro che erano stati uccisi					1	0.71%
coloro che furono uccisi	1	0.75%				
coloro che sono stati uccisi	1	0.75%				
folla	4	3.01%				
gente	3	2.26%			1	0.71%
individui					2	1.43%
Maidan	1	0.75%				
manifestanti					1	0.71%
morti	1	0.75%				
persone	104	79.20%	6	85.71%	114	81.43%
persone uccise	7	5.26%	1	14.29%	16	11.43%
presenti					1	0.71%
protestanti					1	0.71%
tutti coloro che erano stati uccisi	2	1,50%				
uccisioni					1	0.71%
vittime	1	0.75%			1	0.71%
vittime uccise	6	4.51%				
Totals	133	100%	7	100%	140	100%

From Table 4, it is again evident that there is less variety in the solutions the post-editors chose, although perhaps a little less so. However, the difference is again statistically significant, as can be verified using the contingency table below (Table 5).

Table 5. Lexical variety contingency table for *people*

	Translation	MTPE
persone	104	114
persone uccise	7	16
Other n-gram	22	10
(χ^2 (2, N = 273) = 8.31, p < .05).		

The method described above of dividing the texts into two sets according to the number of occurrences of *persone* results in six misattributed texts. In other words, the human translations and MTPE may be identified with $14/20 = 70.00\%$ accuracy.

3.4 Degree of naturally occurring lexical variety in DeepL Translator raw output

To measure the degree of lexical variety naturally produced by DeepL Translator in its raw output, two longer texts were machine-translated, containing a number of MT markers equal to the number of students involved times the number of MT markers

found in each of the two shorter texts translated/post-edited in the main experiment (18 x 4 = 72 for *there are*; 19 x 6 = 114 for *people*). These longer texts were put together by taking whole paragraphs rich in the n-gram concerned from several Wikipedia articles and pasting them all into a single document. The raw MT output from this experiment was found to be less lexically impoverished than in the equivalent experiment reported in 2018, at least as regards the two n-grams studied. In the case of the first MT marker considered (*there are*), the number of translation solutions in the raw output from DeepL Translator (8) is quite a lot smaller than the number used by the human translators (14), and the distribution of the HT solutions is more even. However, the most chosen solution (*ci sono*²), had exactly the same frequency in the HT and the raw output (50%). In the similar experiment reported in 2018, DeepL Translator had translated *there are* with *ci sono* 90% of the time.

Regarding the second MT marker (*people*, as the plural of *person*), the translation solutions in the raw output (11) were only slightly less numerous than those chosen by the human translators (13) but the solutions themselves were often quite different.

Due to the presence of a lot of very low frequency translation solutions and translation solutions occurring only in the HT and not in the raw output and vice versa (zero values), meaningful chi-square statistical analysis is unfortunately not possible.

By way of comparison, the same longer texts were also fed to Google Translate, whose raw output showed much clearer signs of lexical impoverishment (only 3 solutions for the first MT marker and 7 for the second).

3.5 Task questionnaire

The students completed a short questionnaire after they delivered their files. They were first of all asked how they had done the translation. The majority wrote their translations in a new Microsoft Word file (Table 6).

Table 6. How the translation was done in Microsoft Word

	Number of replies
New empty Microsoft Word file	26
Overwrite original text	12
Create two column table	3
Write underneath, then delete original text	1

They were then asked what reference material they had used while translating or post-editing (Table 7).

² Variants required for grammatical reasons, such as *ci siano* (subjunctive tense), were considered to be the same solution.

Table 7. Use of reference material while translating or post-editing, multiple answers were allowed.

	Translating	Post-editing
Online dictionaries, encyclopaedias or web-sites	42	40
Asked a colleague for help	4	1
Physical, printed reference material	0	0
No reference material	0	2

The results show quite clearly that print dictionaries are a thing of the past.

The students were instructed not to use MT engines in any way to prevent the translators from turning their task into a second post-editing assignment. It is clear from the tables above that the same kinds of materials were used for both processes. Two post-editors did not refer to any external reference material.

The students were asked to assess how useful it was to be able to refer to the source language text while post-editing (8.00 ± 1.89 SD points out of 10) and to the original unedited raw output during the self-revision of their post-editing (6.12 ± 3.05 SD points out of 10). These results substantially confirm the ISO 18587:2017 definition of post-editing as involving three texts: the source text, the MT output and the final target text [8]. They also suggest that monolingual post-editing would have been ill advised in the case of the texts chosen.

Another question the students were asked was if they would have used MT in some way during their task if it had been allowed (Table 8).

Table 8. Number of translators and post-editors who would have used MT if it had been allowed

	Number of replies
Never	5
Only during the post-editing	0
Only during the translation	24
Both during the post-editing and the translation	13

3.6 Syntactic differences

3.6.1 Slovakia text segmentation

There were nine segments in the original text and in the machine translated text before post-editing. Table 9 shows the number of translators and post-editors who either split or joined segments at least once.

Table 9. Number of translators and post-editors who joined or split segments in the Slovakia text

Segments	N. translation (18)	N. post-editing (20)
Split/join	9	7
No split/join	9	13
$(\chi^2 (1, N = 38) = 8.74, p < .05).$		

The difference is not statistically significant. So, the translators and post-editors felt equally free to split/join segments.

3.6.2 Ukraine text segmentation

There were fifteen segments in the original text and in the machine translated text before post-editing. Table 10 shows the number of translators and post-editors who either split or joined segments at least once.

Table 10. Number of translators and post-editors who joined or split segments in the Ukraine text

Segments	n. translation (18)	n. post-editing (20)
Split/join	6	4
No split/join	14	16
$(\chi^2 (1, N = 40) = 0.53, p < .05).$		

The difference is again not statistically significant. So, the translators and post-editors felt equally free to split/join segments.

4 Conclusion

These conclusions are drawn on the basis of two short texts of only one genre, which limits the generality of the findings to some extent. This limitation is however inevitable since the experiment was carried out as part of an undergraduate degree course and only a limited amount of time could be devoted to it.

In the case of the particular texts used in this experiment, the priming received from the raw output led to MTPE that is distinguishable from HT with a success rate of between 70 and 75%, which is however down from the 100% success rate observed in the 2017/2018 experiment. On the basis of these results, we are led to conclude that the lexical impoverishment phenomenon is indeed attenuating with DeepL Translator. It is however apparent that the results would have been different with Google Translate, which produces raw output with clearer lexical impoverishment, as was mentioned in section 3.4 above.

Despite not being given any particular post-editing guidelines, there was an increase in productivity of between 39.50 and 47.99%. Some of the translators and post-editors chose exactly the same translation solutions for the n-grams studied as were found in

the raw output in precisely the same places. Therefore, the solutions in the raw output are acceptable. Consequently, if the post-editors had strictly applied the ISO 18587:2017 post-editing recommendation to use as much of the MT output as possible [8], the post-editors would not have altered these solutions making the MTPE even easier to distinguish from HT. If we can generalize these results, this fact, together with the increase in productivity, suggests that, in the case of texts where lexical uniformity would make the translation less interesting to read and less intellectually stimulating, such as in the fields of marketing, advertising, literature, journalism, education, entertainment, and creative writing in general, it is neither necessary nor advisable to apply this ISO 18587:2017 recommendation.

Another way of avoiding lexical impoverishment may be to avoid MTPE entirely and use MT as a tool during the translation process, for example in one of the various ways that emerged from the previously mentioned survey among professional translators, such as *for inspiration* or *as a dictionary* [7]. However, this would almost certainly not lead to anything like the increase in productivity achieved with MTPE.

The students found it useful to refer to the source language text during the post-editing and to the original unedited raw output during the self-revision of their post-editing. This suggests that monolingual post-editing in the case of the texts chosen would have been ill advised.

No significant difference was found in the number of segments in the target texts. Evidently, the translators and post-editors felt equally free to split and join segments during their task. Obviously, since some of the post-editors chose not to alter the segmentation of the raw output and therefore found it acceptable, if the previously mentioned ISO 18587:2017 recommendation [8] had been strictly applied, none of the post-editors would have changed the segmentation thus making MTPE more distinguishable from HT syntactically. However, it would be interesting to repeat this experiment asking the participants to use a CAT or post-editing tool to see if they feel equally empowered to join and split segments.

References

1. Castilho, S., Resende, N. Post-Editese in Literary Translations. In *Machine Translation for Conquering Language Barriers*, Special Issue (2022).
2. Castilho, S., Resende, N., Mitkov, R: What Influences the Features of Post-editese? A Preliminary Study. In: *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pp. 19–27. Incoma Ltd., Shoumen, Bulgaria.. (2019).
3. Collodi, C. *Le avventure di Pinocchio. Storia di un burattino*. Libreria Editrice Felice Paggi (1883).
4. Collodi, C. *The Adventures of Pinocchio* (Translation by Carol Della Chiesa). The Macmillan Company (1926).
5. Daems, J., De Clercq, O., Macken, L.: Translationese and post-editese: How comparable is comparable quality? *Linguistica Antverpiensia New Series - Themes in Translation Studies* 16:89–103 (2017).
6. Farrell, M.: Machine Translation Markers in Post-Edited Machine Translation Output. In: *Proceedings of the 40th Conference Translating and the Computer*, pp, 50–59. AsLing: The International Association for Advancement in Language Technology (2018).

7. Farrell, M: Do translators use machine translation and if so, how? Results of a survey held among professional translators. Presented at the 44th Conference Translating and the Computer. Preprint of peer-reviewed paper awaiting publication, DOI:10.13140/RG.2.2.33996.69768, (2022).
8. International Organization for Standardization. ISO 18587:2017: Translation services – Post-editing of machine translation output – Requirements (2017).
9. Stangroom, J.: Chi-Square Test Calculator, www.socscistatistics.com/tests/chisquare2/default2.aspx, last accessed 2023/04/23.
10. Stangroom, J.: How to Report a Chi-Square Test Result (APA), www.socscistatistics.com/tutorials/chisquare/default.aspx, last accessed 2023/04/23.
11. Toral, A.: Post-editeuse: an Exacerbated Translationese. In: Proceedings of Machine Translation Summit XVII: Research Track, pp. 273–281. European Association for Machine Translation, Dublin, Ireland (2019).
12. Volkart, L., Bouillon, P: Studying Post-Editese in a Professional Context: A Pilot Study. In: Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, pp. 71–79. European Association for Machine Translation, Ghent, Belgium (2022)

Effective Editing in Respeaking: Unveiling Uniquely Human Skills in Live Speech-to-Text Transformation

Tomasz Korybski¹[0000-0003-2353-0816] and Elena Davitti²[0000-0002-7156-9275]

¹ University of Surrey, Guildford, UK

² University of Surrey, Guildford, UK

t.korybski@surrey.ac.uk

e.davitti@surrey.ac.uk

Abstract. Set in the context of a rapid technological development in translation, interpreting and speech recognition technology, this paper takes a close-up view of an area where human input still seems to be an indispensable and very hard-to-replace: the area of live speech transformation. To contribute to answering inevitable questions about the future role(s) of humans in real-time language-related practices, we focus on effective edition (EE, Romero-Fresco and Pöchhacker 2018) as the uniquely human skill to deviate from the source message in a way that does not cause any loss of information and may even improve the final output. Firstly, we discuss this concept in the context of earlier research on paraphrasing in translation, reduction in respeaking, and reformulation in interpreting. Secondly, we explore this form of language transformation in practices characterised by immediacy in the way content is conveyed, focusing on the spoken-to-written form. To show the predominant types of effective editions (EEs), we use authentic data from human intra- and interlingual respeakers collected during two related research projects conducted at the Centre for Translation Studies (University of Surrey) in the years 2019-2023. The data in the paper show a variety of EEs s live human interventions that can be mostly grouped into forms of condensation, re-expression, and compensation. Our discussion will centre on the diverse form that EEs can take in various practices, and the importance to recognise that they are context-dependent and can entail dynamic content shifts. We will conclude by emphasising the importance of understanding and categorising the range of these human-performed edits for evaluation, training, and the future advancement of automated solutions powered by large language models.

Keywords: Effective Editions, Interpreting, Respeaking, NER Model, NTR Model, Live Speech-to-Text

1 Rationale and Objectives of this Paper

The object of analysis of this paper is the phenomenon of effective editing in the live speech-to-text practice of intra- and interlingual respeaking. Respeaking can be defined as a “technique whereby a respeaker listens to the original sound of a (live) programme or event and respeaks it in [in the same language, or translates it in another language and in real time], including punctuation marks and some specific features for the deaf and hard-of-hearing audience, to a speech recognition software, which turns the recognized utterances into subtitles displayed on the screen with the shortest possible delay” (adapted from Romero-Fresco 2011: 1). This is a challenging task, which requires rapid decision-making to be able to let the original meaning travel through a language (from source to target language) and a diamesic transfer (from spoken input to written output). To this end, respeakers may reformulate in real time, either in the same language (in intralingual respeaking) or in another language (in interlingual respeaking), and this phenomenon is the focus of this paper. Romero-Fresco and Pöchhacker (2017: 159), define **Effective Editions** (EEs) in respeaking as deliberate “deviations from the source text that do not involve a loss of information or that even enhance the communicative effectiveness of the subtitles”. Such ‘deviations’ are representations of respeakers’ flexibility in transforming the source for the audience without a loss of meaning. To contextualise the phenomenon, we notice its affinity with ‘paraphrasing’ and ‘quasi-paraphrasing’ – interventions aimed at tackling the lexical, grammatical, delivery-related, or stylistic challenges posed by source content. EEs are highly context-specific, which results in a set of distinctive features of effective editing that we present in the examples (section 2.3), seeking to highlight the differences between the live speech transformation workflow within one language (intralingual respeaking) and across two languages (interlingual respeaking).

Rather than striving to produce an exhaustive catalogue of editing phenomena in respeaking, our datasets have led us to suggest a broader, tri-partite classification of EEs that may prove useful in evaluation and training. The extensive dataset gathered in one of the projects that have contributed data for this paper, SMART (see section 2.1), has enabled us to identify potential patterns bottom-up and develop a taxonomy from them. Importantly, in SMART we discovered a statistically significant relationship between the number of EEs and performance, indicating that a higher use of EEs leads to higher content accuracy. As a result, we believe there is a need for more comprehensive research into EEs, which could offer valuable insights for training and quality assessment, as well as for informing human-like transformation practices in speech-to-text and speech-to-speech communication.

Furthermore, after reviewing the existing research on live transformation of speech by human professionals working in live assignments (intra- and interlingual respeakers, interpreters), we have identified a gap in the current body of knowledge on the topic that can be (at least partially) filled with the data we have accumulated from two research projects presented in section 2.1 below. Firstly, although researchers have recognised similarities in paraphrasing/transformation activities implemented by respeakers and interpreters, no taxonomy or classification has been created so far to capture what is generically expressed by Romero Fresco’s and Pöchhacker’s (2017) definition.

Following the near-equivalence paradigm and the concept of quasi-paraphrasing, we therefore take a more specific look at effective editing phenomena to classify them in a pragmatic, evidence-based and bottom-up manner. We claim that the notion of effective editing is useful to capture and describe a range of transformation practices applied by respeakers and, by extension, other real-time speech-to-text practices, because it goes beyond the definition of paraphrasing, reformulation and explicitation previously applied in research. Secondly, to add to the existing research on similarities and differences between intralingual and interlingual respeaking, we show how respeakers use effective editing strategies in concrete contexts. Thirdly, we present live editing strategies as a prime example of a (still) uniquely human skill that may pose a challenge for automated solutions, especially in real-time scenarios. We situate the presentation of EEs from our research within the functional paradigm in translation and interpreting studies, specifically Gile's (1990), Seleskovitch's and Lederer's (1975) *theorie du sens* and Vermeer's (1978) *skopos* theory. According to this approach, the language professional's decision-making process in determining what is a *functional* and, therefore, non-omissible component is contextual, and may vary on a moment-by-moment basis, thus impacting how target messages are formulated also in respeaking practice.

1.1 Relevant Theories in Language Transformation

All speakers of a language possess the ability to transform speech and text using a range of methods, which may vary depending on specific communication and situational needs. From basic paraphrasing to advanced summarization, explicitation, or reduction skills, humans transform language in a variety of ways, within and across languages (as in translation, interpreting) as well as modalities (as in intra/interlingual respeaking). Moreover, different modes of content delivery can be used in this process, such as the use of visual representations like pictograms or emojis in place of letters of the alphabet.

The transformation skills discussed here became a focal point for research already in the 1950s, when the term *paraphrasing* was first widely used and researched by Chomsky (1956). Chomsky's Transformational-Generative model of language explained the transformation of language through syntactic operations, such as reordering and substituting words. The notion of a single underlying structure of language that can be subjected to different rules to modify it gained traction and impacted further investigation of how language can be transformed (including early research on Machine Translation, Bar-Hillel 1953). Since then, research into linguistic paraphrasing has expanded, with new theories and methods being developed. An impactful theory of linguistic paraphrasing is the Meaning-Text Theory proposed by Zholkovskiy and Mel'čuk (1965). This theory suggests that meaning is derived from the context of language, with a lot of emphasis on the semantic (lexical) layer which is the foundation for syntactic and morphological processing based on nodes and arguments linking lexemes.

More recently, with the development of machine translation and natural language processing in general, a linguistic analysis of paraphrasing has become an area of interest for computational linguists. In their presentation of paraphrasing types, Bhagat and Hovy (2013) contend that a more elaborate definition of paraphrases is necessary, departing from the generic understanding of paraphrases as utterances that carry the same

meaning as the original utterance but expressed with a different wording. To this end, Bhagat and Hovy build on linguists such as De Beaugrande and Dressler (1981), Mel'chuk (2012) and Clark (1992) to use the notion of *approximate conceptual equivalence* of paraphrases or 'quasi-paraphrases', which moves away from the unrealistic expectation of strict synonymy that may have been inferred from the generic definition. Bhagat and Hovy introduce a set of 25 operations to specify how paraphrasing can be performed – all of which fall under the supra-categories of re-expression and condensation that we propose later in this paper. They argue that although paraphrasing is a complex and multifaceted transformation activity, it can be structured by listing different paraphrasing phenomena in a systematic way. We have adopted a similar approach to categorising EEs but decided to introduce supra-categories including the category of compensation which is intricately linked to live speech transformation and as such is not captured by Bhagat and Hovy's analysis. The value of having a simple taxonomy is that it facilitates bottom-up analyses and may lead to streamlining both qualitative and quantitative analysis of live speech-to-text and speech-to-speech data.

1.2 Effective Editions, Paraphrasing and Strategic Reformulation in Real-Time Language-Related Practices

The classification quoted above is important for our discussion of transformation of live speech as it offers a broad catalogue of phenomena, many of which we have encountered in our data, too. However, it does not capture the specificity of live speech transformation in full – hence this contribution. Before discussing the categories of effective editing from our data, it is important to define the relevant professions and the settings where live transformation of speech takes place, and how different yet related concepts have been defined to refer to relevant phenomena of language transformation.

Simultaneous Interpreting is when a human interpreter listens to and simultaneously interprets the speech into a target language, usually with special equipment (AIIC 2013). Interpreting is referred to in this paper because researchers have underlined its affinity with respeaking (Robert and Remael 2017, Eugeni 2008, Romero-Fresco and Pöchhacker 2018, Davitti & Sandrelli 2020). Drawing up a competence model for respeakers, Pöchhacker and Remael (2019) emphasize the elements that interlingual respeaking shares with interpreting, i.e., the interlingual and real-time nature of the effort, with the added and challenging component of editing. Due to the required immediacy of simultaneous interpreting and respeaking, the decisions of the language professional performing the task must be made with minimum latency, and the transfer between two languages means that complex equivalence challenges often need to be resolved within seconds or even fractions of a second. This promotes paraphrasing as an efficient strategy to satisfy the audience's communication needs while meeting the time pressure involved in this technique. Researchers interested in the process of interpretation, and especially its Production Effort component (Gile 1995) have looked at paraphrasing as an effective and indispensable tool for interpreters. Paraphrasing and strategic reformulation exercises have long been present in the core of most interpreter training curricula (Moser Mercer 1998). Furthermore, Seleskovitch (1978) and other researchers proposed a model of interpreting theory that includes a *deverbalisation* phase between

understanding the source message and re-expressing its content in the target message. This means that content is processed semantically, focusing on its sense rather than on its lexical form. In a related vein, Gumul's (2017) extensive analysis of explicitation techniques in interpreting builds on the work of more than 10 interpreting researchers and offers a detailed breakdown of explicitation-type reformulations into categories which are also highly relevant for the bottom-up analyses we performed in both our projects.

Respeaking. Language transformation in respeaking has been a topic of research for over a decade now (e.g., Romero-Fresco 2011). An important part of the research (and the pragmatic perspective on respeaking as a service with concrete clients and end users) has been the evaluation of respeakers' output. The approach most widely used in the assessment of accuracy of intra- and interlingual respeaking is based on errors (the NER and the NTR models, respectively – see 2.2 below for more details) identified in a close-up linguistic analysis of (fragments) of transcribed source inputs and target outputs. However, the two recognised models for product accuracy assessment include a component of edition evaluation (called correct editions in the NER, and effective editions in the NTR) intended as positive shifts in the target text with respect to their corresponding source. As previously explained, these represents a positive deviation from the source, which are not penalised in the model since they do not alter the meaning or result in any loss of information. In fact, such changes may even improve the effectiveness of the message being conveyed.

Even though paraphrasing has been mentioned as a common denominator for practices where content is transferred across two languages, it is important to emphasize that in the context of written translation the source text is often 'clean' and in its final version. Conversely, source speeches performed in a live settings are frequently abundant in *ad-hoc* formulations, redundancies, disfluencies, overlapping speech, or simply sub-optimal quality of audio due to speaker's pronunciation and background noise, all of which place an additional burden on the human professional's speech transformation. Importantly, this overarching challenge is shared by both interpreters and respeakers, leading to a unique set of transformation phenomena discussed in this paper. Researchers emphasize the similarities between the two professions as regards preparation, processing, and delivery: the 'simultaneity' of reception of audio and own production, balancing the capacity of one's own working memory, processing and output monitoring are indispensable for the service to meet its communication objectives. These similarities have led Chmiel et al. (2018) to conduct a study on paraphrasing skills in interpreters, respeakers and bilinguals where interpreters tended to be better at semantic redundancy elimination and the ensuing production of concise output, although the advantage was not very clearly pronounced (Chmiel et al. 2018). In a study focusing on reduction practices in (intra)lingual respeaking, Luyckx et al. (2010) found that condensations and omissions seem to be a planned process actively shaped by respeakers depending on external factors such as source text speed and availability of 'ommissible' respeaking units. The same study also presents a list of source reduction strategies that enable the respeaker to retain (much of) the original information despite a significantly condensed output. These include some of the strategies that are also included in our

classification in later sections: omissions, shifting questions into affirmatives, and simplifying indicators of modality (*ibid*, pp. 31-33).

2 Effective Editing in Respeaking Datasets – Classification and Comparison

2.1 Sources of Data

For the purpose of this paper, we looked at EEs across the empirical datasets from two distinct projects carried out at the Centre for translation Studies (University of Surrey). Below we provide a brief overview of both projects.

The **SMART** project (**Shaping Multilingual Access through Respeaking Technology**, Economic and Social Research Council UK, ES/T002530/1, 2020-2023) aimed to investigate the emerging technique of *interlingual respeaking* for delivering real-time speech-to-text services across languages. Specifically, the project analysed the interlingual respeaking performances of 51 language professionals from various backgrounds, including consecutive and simultaneous interpreting, written translation, pre-recorded subtitling, and live subtitling. The study aimed to explore the competences underlying such complex process, the level of accuracy achievable by language professionals with at least 2,000 hours of practice in one or more of the disciplines listed previously, and how to optimise upskilling based on empirical research insights. Participants received 25 hours of bespoke training-for-testing course, which exposed them to both intralingual and interlingual respeaking. They were required to speak English paired with Italian, Spanish, and/or French, with at least one of these languages as their mother tongue. They also underwent six tests, including three intralingual and three interlingual ones, designed to expose participants two different scenarios: speed (fast speakers), planned/unplanned delivery (partially improvised, partially prepared speech), and multiple speakers (interview scenario). This paper refers to the interlingual dataset generated by SMART, which comprises 153 performances of 15 to 20 minutes each, totalling over 2,300 minutes of recorded and analysed performance. This dataset provides a rich source of interlingual respeaking data with aligned sources and targets which were evaluated using the NTR model for accuracy evaluation.

The **MATRIC** project (**Machine Translation and Respeaking in Interlingual Communication**, Expanding Excellence in England, 2019-2022) investigated an alternative speech-to-text interlingual communication workflow involving a human intralingual respeaker providing live subtitles in English via speech recognition software (Dragon) and machine translation (the EU's eTranslation) to translate the English subtitles into multiple languages: French, Spanish, Italian, Polish, German, and Romanian. First, an evaluation of intralingual respeaking data from four professional respeakers performing a total of 12 speeches (each professional respoke intralingually three authentic English language speeches from the European Parliament's speech database) enabled us to single out the best performances, which were later machine-translated to produce a total of 18 performances. 12 of these (for Italian, Spanish, French and Polish) were analysed and compared with the corresponding 12 transcripts of actual performances of four interpreter tandems from the European Parliament's booths working at the same events (also available through the EPTV database). Source speech duration

varied between 3 and 14 minutes, representing the typical range of length of EuroParl debate interventions. Additionally, care was taken to select speeches with varied speech types (improvised and partially improvised, varied topics, speakers with different speeds and accents including one native English accent). Overall, approximately 200 minutes of speech were analysed using both the NER and NTR methods, as explained in the next section.

2.2 Capturing Effective Editions – the NER and NTR Models

Both projects focused on accuracy, measured via the NER and NTR models for intralingual and interlingual evaluation respectively. The NER model (Romero-Fresco 2011, Romero-Fresco and Martínez 2015) is a recognised method for evaluating the accuracy of live subtitles produced through *intralingual* respeaking in media or live event broadcasts. The letters in the model’s acronym stand for the total number of words in the live subtitles (N), edition errors (E), and recognition errors (R). The percentage of accurate content is calculated by subtracting the E and R values from the N value, and then dividing it by N. The NTR model (Romero-Fresco and Pöschhacker 2017) is based on the NER model and used to evaluate *interlingual* respeaking data. In the NTR model, translation errors (T) replace the edition errors to account for the accuracy of interlingual transfer. Errors captured with the NTR include omission, addition, and substitution errors (content errors), as well as correctness and style (form errors). Despite a different naming convention in each of the models, the levels of severity in both the NER and the NTR are as follows: minor (0.25 penalty point), standard/major (0.5 penalty point), and serious/critical (1 penalty point). Importantly and crucially for the models’ differentiation from WER-like models, despite the focus on errors, they leave scope for capturing EEs (also referred to as Correct Editions in the NER model). The analysis grid we used was adapted by Davitti and Sandrelli (2020) from the NER score spreadsheet for the evaluation of intralingual respeaking data. The spreadsheet used for NER and NTR evaluation allowed for segmentation and alignment of source and target’s idea units, to enable investigation at a micro level. The grid also features a special column for capturing and commenting on EEs, which makes it possible to count them as well as extract and examine more closely examples of interventions by language professionals. This repository was analysed qualitatively by six researchers with extensive experience in respeaking and interpreting research. Importantly, as the SMART project offered very extensive data, initial characterisation and taxonomy was done on the basis of SMART’s dataset and then applied to MATRIC. As a result of the analysis, we understood that, firstly, it is possible to distinguish predominant categories of EEs, and secondly, that the current definition of EEs could benefit from expanding and specification based on our authentic data. Below we present these categories with brief definitions and examples sourced from our data.

2.3 Classification and Examples of Effective Editions

Upon separate analysis of the datasets for the purpose of their respective projects, we discovered a clear pattern in the EE types identified, which allowed us to categorise them into three distinct supra-categories. The first category consists of non-penalised

omissions, which typically resulted in effective **condensation**, as not all omissions lead to information loss. The second category is **re-expression**, which can be broken down into two major dimensions: lexico-semantic and structural. The third category identified in the data is **compensation**, which occurs when missing information from a previous fragment is compensated in a later fragment. The following examples and extended definitions will provide a more comprehensive explanation of our tripartite classification. Please note that for the purpose of succinctness and consistency in this paper, we focus on English into English intralingual respeaking and Spanish into English / English into Spanish interlingual respeaking. However, we discovered the same patterns across all other language combinations explored in the two projects. Rather than producing an exhaustive catalogue of all possible instances of condensations or re-expressions, in this paper we focus on presenting examples that showcase the general mechanics of effective editing. The sequence of examples in each supra-category is as follows: examples from the intralingual dataset followed by examples from the interlingual dataset. The effectively edited fragments are in bold print. ‘BT’ stands for back-translation (in the interlingual examples).

Condensations

Condensations are the predominant category in the datasets. At *micro* level, they occur when a source idea unit is compressed and expressed in a shorter form in the target idea unit without (any considerable) loss of information. In our data, condensation was implemented primarily through omissions of redundant information, deictic expressions, or grammatical interventions. Condensations at *macro* level occur when a target idea unit can be expressed in a much more concise way by referring to the preceding idea unit or utilizing elements of content from it., and when a target idea unit captures more than one source idea unit. This type of condensation is typically achieved using pronouns to replace names, or deictics such as this, that, these, those, now, then, here.

The examples below show two types of condensation sourced from the datasets. All the source data transcriptions contain all the words that were in fact uttered by the original speaker. As a result, repetitions and redundancies are present in the source transcripts to demonstrate the full scope of the human intervention in the target.

Example 1 (intralingual respeaking)

Source: Despite the the dramatic and I would even say tragic events **we are just going to discuss in a minute** the first time we see each other after the after the Christmas break and **therefore I really** would like to wish all of you **all the citizens you represent and all the European Union** all the best in the New Year and happiest 2020

Target: Despite the dramatic and I would even say tragic events, it's the first time we have seen each other after Christmas and I would like to wish all of you all the best in the New Year and the happiest 2020.

Example 1 was significantly condensed by the respeaker: the original 330 characters with spaces were turned into just 205 characters with spaces in the respoken version. This resulted in shorter, more readable captions, with the message largely intact.

Thanks to the respeaker's intelligent interpretation of the context, they were able to leave out a whole string of words that did not contribute important information, thus streamlining the entire process for themselves and for the recipients. This depth of intervention could not be expected from an ASR solution – although some redundant elements such as the obvious repetitions resulting from hesitation or uncertainty would have been cut out by currently available speech recognition tools.

Example 2 (interlingual respeaking)

Source: **Entiendo que puede** ser algo confuso así que **quizá** debería explicarlo mejor

BT: *I understand that it can be somewhat confusing, so perhaps I should explain it better.*

Target: It might seem confusing, so I should clarify that.

Example 2 features condensation on two levels: firstly, the rapport-building '*I understand*' which has an important function in conversations, but not necessarily in live captions, is dropped. Secondly, '*quiza / perhaps*' is dropped, leaving the transfer of the sentence's modality to the modal verb '*should*'. Arguably, these changes are justified, although the lack of the 'softening' of the sentence's style by means of 'perhaps' results in a more direct affirmative clause. Overall, however, and in the wider context of this example, these modifications bring no detriment to the original's message and result in improved readability. Although there may be contexts where even a slight shift in modality will interfere with the message conveyed (e.g., a legal context), the respeaker's judgment of the situation (often supported by an assignment brief and preparation) should limit editing to the fragments of source text that lend themselves well to such transformation.

The condensations shown above show the uniquely human skill of editing based on split-second decisions and contextual judgment that result in well-readable output either in the same or a different language. Although today's large language models and text summarization solutions cope well with text reduction (leading to condensation), there remain at least two major areas of human superiority. Firstly, apart from redundant content, all content words in automatically summarized fragments of texts tend to stay in the output. Humans, in turn, are able to 'filter out' non-crucial words, including content words, based on the context. This results in succinct and easy-to-read captions. Secondly, human respeakers can switch their condensation practice on and off very dynamically within one assignment, while an automated solution would need to be prompted separately for selected fragments of the source speech to be able to mimic a human respeaker's behaviour. These two challenge areas, of course, come on top of all the existing challenges related to ASR (such as overlapping speech, background noise) providing input for any human-like editing operations in a cascaded system.

Re-expressions

Re-expression is about effectively using the lexico-semantic and structural possibilities afforded by the source content to produce successful renditions suitable for live captions. Re-expression can take the form of semantic, syntactic and stylistic interventions

that do not interfere with the message and provide a more readable output thanks to, for example, voice change, sentence splitting, sentence merging or additions or substitutions that are deemed positive and are, therefore, not penalised in NTR evaluation.

Example 3 (intralingual respeaking)

Source: On Sunday yesterday my colleague Commissioner Lenarcic reached out and spoke to the Australian minister Littleproud **and he reiterated** the Union's readiness to assist Australia in this moment of crisis

Target: On Sunday, yesterday, my colleague Commissioner Lenarcic reached out and spoke to the Australian Minister Littleproud and **stated** the union's readiness to support Australia in the crisis.

Example 3 shows a situation where successful re-expression at a syntactic level (merging two utterances) and lexico-semantic level (verb substitution) leads to slight condensation of the target, too. In fact, these two categories of effective editing often co-occur in our data, although re-expression may also lead to the opposite of condensation, i.e., extension of the target if content explicitation is necessary following the respeaker's assessment of the caption reader's needs. In Example 3, both merging the sentences and verb substitution produce a more succinct target: the subject does not need to be repeated, and the verb '*state*' reads much shorter than '*reiterate*'. The omission of '*this moment*' is also justified by the use of the definite article and the noun '*crisis*', which provide enough clarity given the context.

Example 4 (interlingual respeaking)

Source: Entonces, ¿significa esto que las personas que se dedican al rehabilitado intralingüístico pues que no tienen experiencia como intérpretes pueden dedicarse al rehabilitado interlingüístico? Yo diría que sí, absolutamente

BT: *So, does this mean that people who perform intralingual respeaking, since they do not have experience as interpreters, can still perform interlingual respeaking? I would say yes, absolutely*

Target: So this means that people who are intralingual respeakers who are not interpreters can work as interlingual respeakers if they work hard.

In Example 4 we can see re-expression through two interventions: at a structural level, i.e., the respeaker changed the source's rhetorical interrogative into an affirmative – as the answer was provided by the speaker in the next sentence, the shift was possible; and at a lexico-semantic level, through an addition which is perfectly plausible on the basis of the context preceding this segment and actually clarifies it further. Such processing is evidence of the respeaker's understanding of the speaker's intention, and their ability to operate syntax and form to express meaning to the best of their abilities.

Compensations

Working with live input provides respeakers with unique opportunities to change the sequence of information presentation if such a change is possible, i.e., in situations when chronology of the presentation is not crucial. Compensation consists in providing

missing information from a previous idea unit (which can be either its part or even an entire idea unit) in later target idea units. Although typically compensations span adjacent idea units, it is also possible for language professionals to compensate for content provided a few idea units earlier.

Example 5 (intralingual respeaking)

Source:	Target:
And, of course, we also have the the pension pay gap which is very very serious.	[no rendition]
The current gender pension gap in Europe stands as over double the gender pay gap at 35.7%	The current gender pension gap in Europe is serious and stands at over double the gender pay gap, at 35.7%.

Example 6 (interlingual respeaking)

Source	Target
Entonces, ¿significa esto que las personas que se dedican al reablado intralingüístico pues que no tienen experiencia como intérpretes pueden dedicarse al reablado interlingüístico?	[no rendition]
BT: <i>So, does this mean that people who perform intralingual respeaking, since they do not have experience as interpreters, can still perform interlingual respeaking?</i>	
Yo diría que sí absolutamente	[no rendition]
BT: <i>I would say yes, absolutely</i>	
Yo diría que sí, absolutamente pero necesitas superar algunos obstáculos mentales del síndrome del impostor	I would say that people that haven't worked a simultaneous interpreters can work as interlingual respeakers, but there are obstacles to overcome.
BT: <i>I would say yes, absolutely, but you need to overcome some of the imposter syndrome's mental hurdles</i>	

Examples 5 and 6 show that source information can be compensated if the respeaker's working memory allows it. Please note that a compensation can also include a different editing category such as re-expression – as is the case in Example 6. Importantly, due to its non-chronological nature, compensation is also uniquely human.

2.4 Discussion

The examples above serve to identify the editing strategies typically employed by intra- and interlingual respeakers. Although such strategies can be found 'in isolation', it is natural for them to co-occur within the same idea unit, as in Example 6. Furthermore, the presence of EEs in interlingual respeaking data may be more pronounced than in intralingual respeaking data. This can be mostly ascribed to the language transfer component, which makes a fully verbatim approach difficult to implement.

Based on the quantitative analysis conducted in SMART, we found EEs were a positive predictor of accuracy, thus confirming their impact on improving interlingual re-speaking output, including the readability aspect. This therefore means that learning more about EEs is important for more precise evaluation, goal-oriented training / upskilling of language professionals, and possible future (partial) automation of the task.

3 Conclusions, Impact and Further Research

Impact on Data Evaluation and Training. The examples from re-speaking data we have cited in this paper show that live spoken text can be subject to a varied range of transformations that we have classified into three macro-categories as condensations, re-expressions and compensations. The extent to which these transformations are applied can depend on many factors, including the briefing of the respeaker, the features of the original speech including speed, structural complexity, genre, and, last but not least, the professional's individual inclinations that may result from one's idiosyncratic disposition, experience and/or training. Our awareness of the range of effective transformations that are applied by respeakers can impact how these professionals are trained. The three broad categories we have proposed may prove useful for providers in identifying any existing strengths and possible training needs for language professionals, and the range and modulation of shifts that can be suitable in a specific scenario. Furthermore, a deeper understanding of EEs can contribute to a more comprehensive and objective application of the NER and NTR models: for example, in some live settings the positive impact of successfully applied EEs may offset some shortcomings of the target (such as minor content omissions). In the near future we plan an experimental re-calculation of the NTR scores from the SMART project including a positive value for EEs to gauge their impact on overall scoring. In the context of education, although strategic reformulation exercises have been part and parcel of the training of various language professionals for decades now, there appears to be the need to study them further in relation to hybrid practices. This is particularly relevant as traditional disciplinary boundaries are blurring, with multimodal practices (like re-speaking) emerging and requiring further investigation. Moreover, there is a need to develop a deeper understanding of the frequency of occurrence of each of EE categories to decide what to prioritise in training. Existing and future data sets can be analysed quantitatively and additional variables such as the impact of source genre on effective editing will need to be taken into consideration when shaping future training courses.

Impact on Resource Development and Automation. Apart from training and evaluation, the classification we have proposed may be used to showcase any shortcomings of existing prototypes in machine interpreting and ASR-based live captioning services. The recent rapid progress in large language models such as GPT 4 also offers promising potential for future development of automatic human-like editing of live speech for the purpose of live subtitling or even interpreting. However, to effectively teach the model and expect consistently good outcomes, we first must understand the phenomenon

thoroughly and collect sufficient data. What we need as a next step is to create larger databases of EEs across different live language-related practices to gain a better quantitative and qualitative understanding of the phenomenon. This will then enable us to identify best practices from human data to train automated large language models.

References

1. Bar-Hillel, Y. (1953). Some linguistic problems connected with machine translation. *Philosophy of Science* 20 (3):217-225 (1953).
2. Bhagat, R., Hovy, E. (2013). What is a Paraphrase? *Computational Linguistics* (2013) 39 (3): 463–472.
3. Chmiel, A., Lijewska, A., Szarkowska, A., Dutka, L. (2017) Paraphrasing in respeaking – comparing linguistic competence of interpreters, translators and bilinguals. *Perspectives. Studies in Translation Theory and Practice*, 26(5), 725-744.
4. Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 113-124, Sept. 1956, doi: 10.1109/TIT.1956.1056813.
5. Clark, E. V. (1992). Conventionality and contrasts: Pragmatic principles with lexical consequences. In: Lehrer, A. and Feder Kittay, E. (eds.), *Frame, Fields, and Contrasts: New Essays in Semantic Lexical Organization*. Lawrence Erlbaum Associates, Hillsdale, NJ, 171–188.
6. Davitti, E., & Sandrelli, A. (2020). Embracing the Complexity: A Pilot Study on Interlingual Respeaking. *Journal of Audiovisual Translation*, 3(2), 103–139.
7. De Beaugrande, R. and Dressler, W.V. (1981). *Introduction to Text Linguistics*. Longman, New York, NY.
8. Eugeni, C. (2008b). A sociolinguistic approach to real-time subtitling: Respeaking vs. shadowing and simultaneous interpreting. *English in International Deaf Communication*, 72, 357-382.
9. Gile, D. (1990). Scientific Research vs. Personal Theories in the Investigation of Interpretation. In: L. Gran & Ch. Taylor (eds.) *Aspects of Applied and Experimental Research on Conference Interpretation*. Udine: Campanotto, 28-41.
10. Gumul, E. (2017) *Explicitation in Simultaneous Interpreting. A study into explicating behaviour of trainee interpreters*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.
11. Korybski, T. Davitti, E., Orăsan, C., Braun, S. A Semi-Automated Live Interlingual Communication Workflow Featuring Intralingual Respeaking: Evaluation and Benchmarking. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pp. 4405–4413 Marseille, European Language Resources Association (ELRA).
12. Mel'čuk, I., (2012). *Semantics: From Meaning to Text*. John Benjamins Publishing Co., Philadelphia, PA.
13. Mel'čuk, I. (2001). *Communicative organization in natural language: The semantic-communicative structure of sentences*. Amsterdam: John Benjamins.
14. Luyckx, B., Delbeke, T., Van Waes, L., Leijten, M., & Remael, A. (2010). Live subtitling with speech recognition causes and consequences of text reduction. *Across Languages and Cultures*, 14, 15-46.
15. Slocum, J. (1985). A survey of machine translation: its history, current status, and future prospects. *Computational Linguistics*, Volume 11 (1), January-March 1985, retrieved from <https://aclanthology.org/J85-1001.pdf> on 13 April 2023.

16. Pöchhacker, F., Remael, A. (2019). New efforts?: A competence-oriented task analysis of interlingual live subtitling. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 18, 130–143.
17. Reiss, K., Vermeer, H. J. (1984): *Grundlegung einer allgemeinen Translationstheorie*. *Linguistische Arbeiten* 147. Tübingen: Niemeyer.
18. Romero-Fresco, P. (2011). *Subtitling through speech recognition: Respeaking*. Manchester: St. Jerome.
19. Seleskovitch, D., Lederer, M. (1975). *Pedagogie Raisonnee De L'interpretation (Traductologie)* (Volume 4, French Edition), 41-44.
20. Seleskovitch, D. (1975): *Langage, langues et mémoires. Etude de la prise de notes en interprétation Consécutive*, Minard, Paris.
21. Stinson, M. S. (2015). Speech-to-text interpreting. In F. Pöchhacker (ed.), *Routledge encyclopedia of interpreting studies* (399–400). London: Routledge.
22. Vermeer, Hans Josef (1978). "Ein Rahmen für eine allgemeine Translationstheorie". *Lebende Sprachen*. 23 (3). doi:10.1515/les.1978.23.3.99. ISSN 0023-9909. S2CID 62754751.
23. Žolkovskij, A., Mel'čuk, I. (1965). "O vozmožnom metode i instrumentach semantičeskogo sinteza (On a possible method and instruments for semantic synthesis)". *Naučno-texničeskaja Informacija*. 5, 23–28.

InterpreTutor: Using Large Language Models for Interpreter Assessment

Cihan ÜNLÜ¹[0000-0001-7857-2031]

¹ Istanbul Yeni Yüzyıl University, Istanbul, Türkiye
cihan.unlu@yeniuyuzuil.edu.tr

Abstract. The recent development of large language models (LLMs) has shown remarkable natural language processing capabilities and created new possibilities for applications in various fields. With their advanced NLP features, such as text comparison, semantic analysis, text summarization, text classification, and text completion, this study aims to investigate whether these models have the potential to be used as tools for evaluating translation/interpretation output in a textual representation. In this paper, we question the capabilities of an LLM and propose InterpreTutor¹, an LLM-powered application with simple UI that utilizes Generative Pre-trained Transformer (GPT-3.5 turbo and GPT-4 models) and speech recognition (OpenAI Whisper) to offer detailed feedback on interpreters' performances based on automatic analysis of the transcriptions of their interpreting practice. The tool's primary focus is to act as an easy-to-use self-tutoring tool, offering feedback based on four human evaluation criteria. While InterpreTutor may not cover all aspects of interpreting performance, such as prosodic features of the delivery, it still provides insights into aspects that can be assessed through textual representation. In this paper, we will discuss the potential of LLMs, the development of InterpreTutor, its underlying methodology, and provide examples of its application with a small-scale experiment.

Keywords: large language models, interpreter training, artificial intelligence, speech recognition, automatic interpreter assessment, gpt-3.5, gpt-4

1 Introduction

The rise of large language models (LLMs) and LLM-powered applications, such as ChatGPT² and DALL-E 2³ by OpenAI, Claude⁴ by Anthropic, PaLM⁵ by Google, LLaMA⁶ by Meta, have brought about new avenues for the full and semi-automatization of the complex natural language tasks. The wide-ranging impacts of

¹ Publicly available at <https://www.interpretutor.app>

² <https://openai.com/blog/chatgpt>

³ <https://openai.com/product/dall-e-2>

⁴ <https://www.anthropic.com/index/introducing-claude>

⁵ <https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html>

⁶ <https://ai.facebook.com/blog/large-language-model-llama-meta-ai/>

LLM and their advanced natural language processing (NLP) capabilities hold significant promise for future derivative LLM-powered end-user applications like support agent assist, personalized tutoring, grammar correction and style, frontend/website generation, database query optimization, code generation and autocomplete, copywriting, etc. Github Copilot, Bing Search, ZoomIQ, Khanmigo, Duolingo Max, In-worldAI, and many other end-user software are commercial examples on the market. The field of translation and interpreting is set to experience such transformative advancements, too. Currently, such advancements are limited to possible benefits such as document-level machine translation (MT) [1], producing synthetic data for training better MT engines, term extraction, glossary creation, and conducting quality estimation. Still, LLMs have the potential to be an integral part of the hybrid workflow of language service providers soon.

This study aims to investigate the capabilities of two of the most advanced LLMs of Open AI: GPT-3.5 and GPT-4, which are the base models used in the publicly available ChatGPT⁷, with a particular focus on interpreter assessment. To see the generative features of an LLM in interpreting evaluation, we introduce an LLM-powered application *InterpreTutor* designed to transcribe the audio recordings and offer comprehensive feedback for performances based on various human assessment criteria. Details and usage scenarios for LLMs in the literature are briefly discussed in 1.1. The architecture of InterpreTutor is outlined in section two. On the other hand, section three presents the methodology of the small-scale experiment comparing the InterpreTutor's performance in GPT-3.5 and GPT-4. Further, an experimental setup was designed to include two trainee interpreters whose performances were recorded in audio and used as input for our tool. InterpreTutor's automatic performance analysis was conducted by leveraging GPT-3.5 and GPT-4. In its analysis, InterpreTutor gave scorings for each of the four criteria. Afterward, the scorings were compared to human evaluators' scorings. We aimed to understand its effectiveness in a restricted task: understanding the context, feedback generation on the completeness of a text, evaluating coherence, handling terminological consistency across texts, and providing an accurate report.

1.1 Large Language Models

LLMs refer to large-scale deep learning models trained on massive amounts of text data to perform various natural language processing tasks. Among these models, the Generative Pre-trained Transformer (GPT) models [2][3] have gained considerable attention due to their ability to generate coherent and context-aware text. Two of the widely known LLMs, GPT-3.5 and GPT-4 are the third and fourth iteration, respectively, in the GPT series of US-based AI research laboratory OpenAI. As a deep learning-based language model, LLMs generate human-like text using an autoregressive approach. In simpler terms, it's a system that creates persuasive, semantically coherent words, code, or data sequences based on a given input, known as the prompt.

⁷ <https://openai.com/blog/introducing-chatgpt-and-whisper-apis>

GPTs, with their traits of general-purpose technologies, have not only opened up new possibilities for advanced natural language tasks but also raised concerns regarding their broader implications, including the economy, social dynamics, and policy development. In terms of the economy, a recent study reported that the rise of LLMs could affect a substantial portion of the U.S. workforce, with around 80% of workers seeing at least 10% of their tasks influenced by these technologies and approximately 19% experiencing impacts on at least 50% of their tasks [4]. In politics, governments are grappling with the rapid advancement of AI and call for regulation, as large language models enable AI to create realistic art, write essays, and generate code within seconds. In education, institutions have started taking measures to address the potential misuse due to concerns about students using it for assignments and the potential for unreliable or inappropriate content [5]. On the other hand, LLMs have a positive potential in creating educational content, enhancing student engagement, and personalizing learning experiences.

The works evaluating the performance of GPT and other LLM models indicate outperforming results in some domains compared to state-of-the-art NLP methods. For instance, GPT-3.5 has shown remarkable performance in text classification [6][7] and has proven to be more precise in MT evaluation compared to statistical metrics [8].

Among many capabilities, GPTs can be employed for the following linguistic tasks as they generate promising results:

- **Machine translation:** GPTs are capable of producing fluent and competitive machine translation outputs that perform closer to state-of-the-art neural machine translation models (e.g., Google) on high-resource European Languages [9][10][11]. However, there is the risk of generating hallucinations for non-English-centric MT tasks [12].
- **Text summarization:** LLMs can generate concise summaries of long pieces of text by understanding the main points and condensing the information into a shorter version [7] [13] [14].
- **Text generation:** LLMs can generate coherent, context-aware, and human-like text in various formats, such as stories, articles, or conversations for any given purpose. GPTs can adapt their text generation to specific writing styles, formats, or tones, allowing for tailored content production based on user preferences [15].
- **Dependency parsing:** LLMs can analyze the grammatical structure of a sentence and determine the relationships between words and phrases.
- **Textual similarity analysis:** GPTs can understand complex semantic patterns and relationships between words and phrases [16].

The list may be extended further for other specific linguistic tasks, such as named entity recognition, part-of-speech tagging, sentiment analysis, etc. The textual analysis and context understanding capabilities of GPTs make them particularly relevant for our research, where we are focused on their performance for analysing the interpreting output. In this context, this study has the following goals:

- a. to explore and evaluate the potential of an LLM (GPT-3.5 / GPT-4) in assessing interpreting performance through zero-shot prompting based on human evaluation criteria while demonstrating the accuracy, comprehensiveness, and constructiveness of the feedback generated.
- b. to launch a system that will provide an accessible tool for trainees that could be used as supplementary feedback in deliberate practice.

1.2 Automatic Evaluation and Scoring in Interpreter Assessment

Testing and assessing spoken-language interpreting has been gaining attention over the past decade, particularly in the fields of interpreter education, professional certification, and interpreting research. Han [17] outlines the assessment procedures for interpreting in three major aspects: (1) specificities of interpreting assessment (modes and directionality), (2) assessment design (difficulty, variety, number, and length of tasks and evaluation criteria), and (3) scoring and rater training. For scoring, different scoring methods are drafted by scholars, like the error-analytic method [18][19], questionnaire-based scoring [20], rubric scoring [21], and comparative judgment [22]. Based on these established views on assessment and scoring, automatic evaluation of spoken-language interpreting focuses on predicting human-assigned scores for interpreting performance. Automatic assessment and scoring techniques embark on a method that aims to minimize human involvement in analysing, assessing, and quantifying interpreting quality. It basically employs specifically designed algorithms to automatically calculate metrics and indices related to interpretation performance. While this approach holds promise in transforming the evaluation of interpreting quality, it remains in its nascent stage and necessitates further comprehensive research [23]. Research-wise, it has gained traction as an important area of research, with numerous studies exploring various methods for predicting interpreting quality. In general, there are three strands of automatic evaluation research [17] (a) researchers evaluate speech fluency using objectively measured temporal variables [24] [25]; (b) analyse linguistic features using corpus linguistic tools [26] [27]; (c) use algorithmic quality evaluation metrics that are already deployed for MT [23] [28].

In the first strand of research, researchers generally focus on analysing the temporal variables relating to fluency [29], like speech rate, phonation time ratio, and mean length of the run, which was proven to correlate with human rater's judgment [24] [25]. On the other hand, the second strand, led by interpreting studies, computer science, and natural language processing researchers, utilizes large corpora and sophisticated machine-learning techniques to analyse diverse surface features and predict human assessment results. For instance, Le et al. [30] built an interpreting corpus of 6,700 utterances and employed word confidence estimation systems integrating automatic speech recognition and machine translation features. The results of the experiments indicate that MT features have the most significant impact on the quality assessment of speech translation, while ASR features provide valuable supplementary

information. Meanwhile, Stewart et al. [31] enhanced the Quest++ quality estimation model with four interpreting-specific features to assess interpreting quality. Li et al., [32] propose a neural-based automatic scoring model for Chinese-English interpretation using a multi-indicator assessment. Three improved attention-based BiLSTM neural models are developed to learn the text of transcribed responses in terms of keywords, content, and grammar. BERT pre-training technique is used for keyword and content vectorization, while grammar is initialized randomly. Additionally, fluency is assessed through speech tempo analysis. An integrated score is derived by combining the four metrics using a random forest regression approach. Experimental results demonstrate that the proposed scoring method is effective and performs comparably to human assessment (manual scoring).

Using MT evaluation metrics for automatic assessment of interpreting constitutes the third strand of research where researchers adapt the translation-based MT evaluation metrics into spoken-language interpreting. For instance, Han and Lu [28] investigated the relationship between machine translation metrics, such as BLEU, TER, BERT, NIST and METEOR, and human scores for interpreting assessment, finding moderate-to-strong correlations that suggest the potential for automation. The study was conducted for English-to-Chinese interpreting tasks within language learning classrooms. The results indicated strong correlations, particularly for METEOR, with sentence-level evaluation showing closer correlations than text-level assessments [28]. Similarly, Lu and Han [23] explored the relationship between automated metric scores and human-assigned scores in assessing English-Chinese bidirectional interpretations performed by 56 students. Five machine translation metrics were used alongside various rater types and scoring methods, such as analytic and holistic rubric scoring and ranking. The study observed a relatively strong correlation between automated metric scores and human scores when BLEU, NIST, and METEOR are in question. These studies demonstrate the possibilities of using machine translation metrics and natural language processing techniques for automatic interpreting assessment.

Indeed, an automatic assessment of an interpreted text using the transcripts only would pose a challenge and be bound to some limitations. Being on a textual level, it will not be able to consider non-verbal cues, fluency, pace, and prosodic features (intonation, stress, rhythm) of the delivery, which are essential components of interpreting assessment. Other non-textual assessment units can be voice modulation, the skill of maintaining calm under pressure, memory retention, etc. In general, a comparative assessment based on source and target text would only provide an overall quality evaluation. GPT models stand out with their ability to handle many NLP tasks which can generate practical output with zero-shot prompting. By making the most of the generative capabilities of GPT-3.5 and 4, we attempt to investigate whether integrating LLMs into the assessment process could provide more accurate and nuanced evaluations, as they can capture linguistic features and semantic understanding beyond traditional metrics. The prompting and evaluation criteria that we embark upon in GPT-3.5 and GPT-4 are outlined in section 2.3.

2 Architecture and Workflow

In this section, we will detail the design principles and objectives behind Interpreter, including the rationale for its focus on textual feedback and the considerations made to ensure its usability and effectiveness in interpreting assessment scenarios.

2.1 User Interface

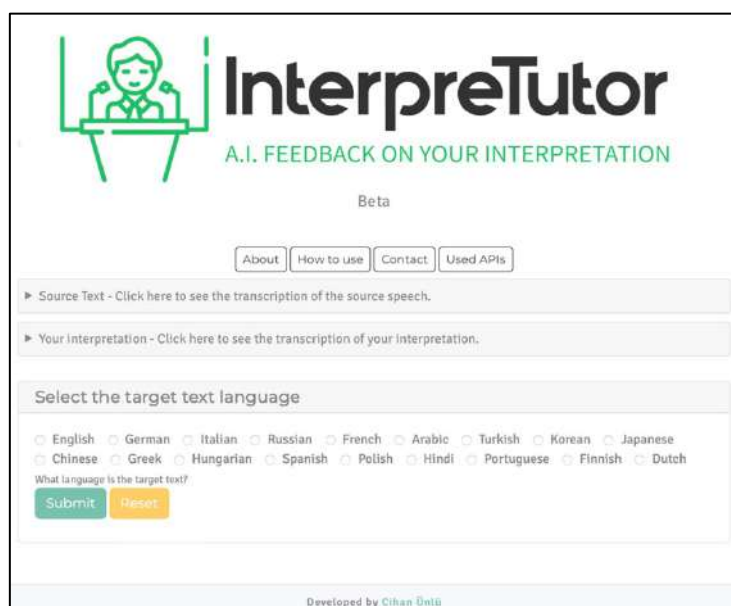


Fig. 1. A screenshot from the main interface of Interpreter

The user interface (UI) of the tool provides users with a step-by-step process to facilitate the evaluation of the given inputs. The user begins by uploading the source speech, which is then transcribed using OpenAI Whisper's Automatic Speech Recognition (ASR) system. Upon completion of the transcription, the user is prompted to select the source language. Subsequently, the user is required to upload their interpretation, which is again transcribed into target text using Whisper ASR API. The user must then choose the target language (as shown in Fig. 1)⁸. To generate a reference text in the target language, the user is given the option to either upload the source speech once more, which is then translated automatically via the Whisper Translate API⁹, or manually input a pre-determined reference text, also known as the golden standard. With all inputs in place, the source and target texts, as well as the reference

⁸ Language selections are done after the transcription/translation processes.

⁹ Whisper's speech translation engine currently only works for translations into English. For other directions, the user is asked to enter the reference text manually.

text, are combined to form the general prompt for the model. The evaluation process commences, and upon completion, the user is presented with a concise evaluation summary, offering insights into their interpreting performance based on the predetermined evaluation criteria (see 2.3.).

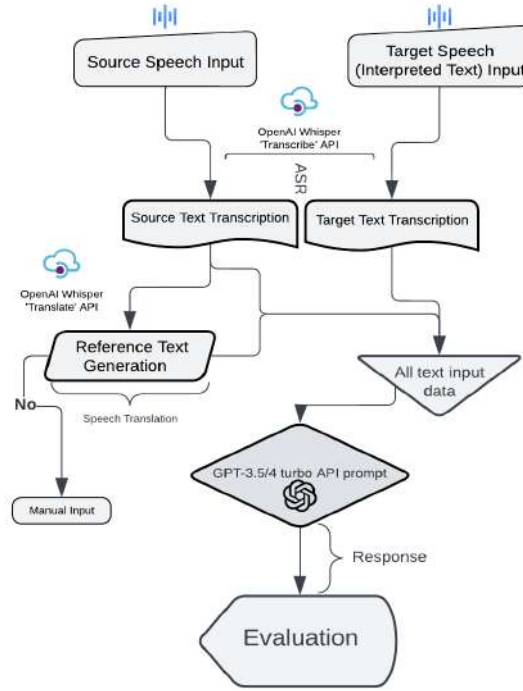


Fig. 2. The workflow applied in InterpreTutor.

2.2 Models Used

In the development of InterpreTutor, we employed two state-of-the-art AI models to facilitate the conversion of speech into text and the generation of assessments, respectively. Both models have been developed by OpenAI and have demonstrated impressive performance in their respective domains.

OpenAI Whisper.

Whisper [33] is an Automatic Speech Recognition (ASR) system that has been trained on a large dataset of multilingual and multitask supervised data collected from the web. Its primary function within InterpreTutor is to convert both the source speech (source text) and the user's spoken input (interpretation) into text to be processed by the LLM. Whisper is an example of the application of transformer models in speech processing is a versatile model designed for speech recognition in challenging environments, such as noisy or low-resource settings. It is capable of multilingual

speech recognition, speech translation, and language identification. In InterpreTutor, we use Whisper API for both transcription and translation. The transcription function is utilized for both source text capturing and target text generation (the user's interpretation), whereas the speech translation (ST) function of Whisper is used for the automatic generation of the reference text. The application relies on the accuracy of Whisper's speech-to-text conversion, which gives promising results with a low word-error rate (WER), especially for the languages English, Italian, German, and Spanish [34]. One thing that Whisper can do is annotate the filled pauses. It is possible to use a prompt to improve the quality of the transcripts generated by the Whisper API. The model may leave out or include common filler words in the audio. We created an example prompt to keep the filler words (e.g., 'uhm(s)' 'hmm' etc.) in the transcript, thereby evaluation of the disfluency is facilitated when the LLM analyses the occurrences of filled pauses.

OpenAI GPT-3.5 & GPT-4.

OpenAI's GPT is a large-scale language model known for its exceptional natural language processing capabilities. In InterpreTutor, we specifically used the model gpt-3.5-turbo and GPT-4¹⁰ (latest iteration) from the GPT family to generate responses and feedback on the user's interpreting performance. Based on the textual input received from the OpenAI Whisper API, GPT-3.5 and GPT-4 are responsible for analysing the transcriptions, comparing the source and target texts, and providing feedback on the quality of the interpretation based on the criteria outlined in section 2.3.

2.3 Evaluation Criteria and Prompt Generation

This section will explain the mechanisms by which InterpreTutor prompts the GPT-3.5 and GPT-4 models to generate feedback, as well as the evaluation criteria used to assess the interpreters' performance. It will also cover the development of these criteria and the considerations made in order to ensure their relevance and fairness.

Evaluation Criteria.

In evaluation criteria, various quality criteria have emerged in the literature with different viewpoints. In their research, Schjoldager [20] identifies four primary factors to evaluate student interpretation performance, encompassing aspects such as comprehensibility, linguistic accuracy, coherence, and fidelity to the source. On the other hand, Riccardi [35] delves into a more granular assessment approach by proposing 17 specific criteria, which cover areas like phonetic and prosodic accuracy, speech production, non-verbal communication, and lexical variations for both general and specialized vocabulary. The key criteria identified by subsequent research include content, delivery, and language quality as the three primary components of interpreting quality [21] [36]. Drawing on Lee's [21] framework for assessing interpreting perfor-

¹⁰ For the technical report, see <https://arxiv.org/abs/2303.08774>.

mance in the classroom, we adopted a similar three-dimensional approach to interpreting quality, incorporating additional sub-criteria. Our evaluation framework includes 1) accuracy, fidelity and completeness (content); 2) textual integrity, focusing on cohesion and coherence; 3) appropriate usage of terminology; and 4) the presence or absence of disfluency markers (only repetitions and disfluency markers).

Prompting.

The GPT-3.5 and GPT-4 API requests two prompts to generate the answer: system and the user. The system prompt sets the context and the behaviour of the assistant by providing specific information or defining the character of the AI. In our case it is: "You are a conference interpreter trainer who makes detailed and objective evaluations on spoken-language interpreting". The user prompt on the other hand is the main instruction that the GPT handles. Our prompt is designed to encompass three essential criteria as specified above: accuracy, fidelity, and completeness; cohesion and coherence; terminology; disfluency markers. Specifically, we create four responsibilities for the language model and draft the prompt accordingly. 1) Comparing the interpreted content with the source content to check for correctness and completeness. 2) Analysing the logical organization and connectedness of the interpreted content, including transitions. 3) Evaluating the terminological accuracy and consistency. 4) Detecting disfluency markers: repetitions and filled pauses. In the prompt, we elaborated on the aspects and the model is guided accordingly. Consequently, the prompt asks the following questions for the criteria and includes the instructions accordingly:

Accuracy & Fidelity & Completeness:

- **a)** Does the target text accurately convey the meaning of the source text in general?
- **b)** Are there any deviations, such as omissions, additions and unjustifiable changes or misinterpretations of the meaning and intention of the speaker of the source text? Are any important details omitted? Are there grammatical mistakes? Specify in detail.

Cohesion and coherence:

- **a)** Is the target text logically connected and coherent in general?
- **b)** Does the text make sense as a whole in general and make a narrative in the language of the target text despite the minor errors? Specify in detail and give examples if there are any coherence and fragmentation issues in this regard.

Terminology:

- Was correct and consistent terminology used in general? If there are any misuses, point out all major errors, and elaborate.

Disfluency markers:

- Detect the occurrences of repetitions (like saying the same word repeatedly and unnecessarily) and filled pauses.

When all inputs are received for the creation of the prompt, the prompt then guides the model to get the assessment based on these criteria. The model is given the task without any prior examples or training data. Therefore, the model is asked to com-

plete the task using only its pre-existing knowledge, without any fine-tuning or contextual clues¹¹.

2.4 Output Example

Based on the evaluation criteria and the tasks given in the prompt, the tool generates the general evaluation. An exemplary short and concise output using GPT-3.5 (without scorings) is as follows:

<p>1) Accuracy & Fidelity & Completeness: The interpreter's translation is mostly accurate and faithful to the source text. However, there are a few important details that have been omitted or inaccurately translated. For example, the interpreter states that the global GDP will decline by 6.6% in 2020, while the source text says that the global non-profit foreign exchange rate will fall by 4.4% in 2020. The interpreter also mentions a 60.2% decline in global trade, while the source text states a 13.2% decline.</p> <p>2) Coherence: The interpreter's translation is mostly coherent but there are some issues with the logical and semantic connection of sentences. For example, the interpreter states that the decline in global GDP is the largest annual decline since the Great Depression of 1931, but in fact, the Great Depression started in 1929. The interpreter also mentions that the pandemic has had a significant impact on the market, but it is not clear which market is being referred to. Additionally, there are some abrupt endings and beginnings in the interpreter's translation, which make it less coherent.</p> <p>3) Terminology: The interpreter mostly uses correct and consistent terminology, but there are some misusages. For example, the interpreter uses "Earth Commercial Corporation" instead of "World Trade Organization", and "employment rate" instead of "unemployment rate".</p> <p>4. Disfluency markers: The interpreter has used some repetitions and filled pauses in the interpretation. For example, the interpreter has repeated the phrase "organizasyon komitesi" unnecessarily. The interpreter has also used some filled pauses such as "söyle" and "yani" which are not necessary.</p>
--

Fig. 3. An example output (Whisper+GPT-3.5) in InterpreTutor

Due to a larger model size than GPT-3, GPT-4 demonstrates superior comprehension and reasoning capabilities¹², which was also observed while experimenting with them in InterpreTutor. Its expanded token limit of 32,000¹³, equating to approximately 25,000 words, notably enhances report generation in our tool, providing more extensive and detailed outputs. While utilizing GPT-3.5, text inputs through ASR and reference text make the prompt too long, giving little token size for the answer generation. However, the max token size attributed in the model reinforces the advantage of employing GPT-4, which is set to 4000. Figure 4 shows another example output but only for the first criteria: Accuracy, Fidelity, and Completeness. In addition to model comparison based on size and capabilities, this study also aims to compare the relevancy of the scorings of each model to that of human judgement. However, it's important to note that the focus here is primarily on a score-based comparison. Despite this limitation, GPT-4 generates more detailed and accurate feedback points, which underlines its potential superiority over GPT-3 in the context of producing logical and accurate output. Yet, both models carry the risk of hallucinations and inaccurate information. They can also produce certain biases inherent in the training data, which can possibly affect the feedback provided. In order to mitigate this, a low temperature (0.1) was set to make the model more deterministic and less random.

¹¹ Apart from the primary system prompt which includes "You are a conference interpreter trainer who gives detailed feedbacks and evaluations on spoken-language interpreting."

¹² <https://openai.com/research/gpt-4>

¹³ This marks a substantial increase from GPT-3.5's 4,000-token limit, equivalent to around 3,125 words.

1. Accuracy & Fidelity & Completeness:

a) The interpreter's translation does convey the general meaning of the source text, but there are some inaccuracies and misinterpretations. For example, the interpreter's explanation of the exchange rate is not as clear and detailed as in the source text. Additionally, the interpreter incorrectly states that the value of the dollar decreases when people sell it, whereas the source text explains that the value decreases when people convert their dollars into Turkish liras. The interpreter also omits the explanation of how exporting and importing affect the currency exchange rate.

b) There are several omissions and deviations in the interpreter's translation. The interpreter omits the explanation of how exporting and importing affect the currency exchange rate, as well as the detailed explanation of how the Central Bank's interest rates affect the value of the dollar. The interpreter also deviates from the source text by stating that the value of the dollar decreases when people sell it, which is not accurate.

Examples of major inaccuracies and omissions:

- The explanation of the exchange rate is not as clear and detailed as in the source text.
- The interpreter incorrectly states that the value of the dollar decreases when people sell it.
- The interpreter omits the explanation of how exporting and importing affect the currency exchange rate.
- The interpreter omits the detailed explanation of how the Central Bank's interest rates affect the value of the dollar.

Rating Score: 6/10

The interpreter's translation does convey the general meaning of the source text, but there are several inaccuracies, misinterpretations, and omissions. The interpreter should work on improving their understanding of the source text and providing a more accurate and complete translation.

Fig. 4. An example output (Whisper+GPT-4) for the content accuracy and completeness generated in InterpreTutor

3 Pilot test

3.1 Methodology

Our methodology for the pilot test involves a speech to be interpreted by two interpreters and later to be evaluated by InterpreTutor and human evaluators. A number-dense speech (duration: 03:01) was chosen to be interpreted from English into Turkish in consecutive mode by two trainee interpreters. The topic of the speech is FIFA World Cup and it consists of 17 named entities and 15 numerical items which makes it challenging and prone to omissions and/or substitutions in interpreting. The source speech and the record of the interpretation (in mp3/wav/mp4 formats) were used as inputs for our LLM-powered tool InterpreTutor utilizing both GPT-3.5 and GPT-4 separately. Apart from the specific instructions in the prompting, the LLM models were instructed to give a score out of ten for each criterion. For the human judgement, both source and target (interpretation) data were transcribed through Whisper ASR (word-error-rate: 0% in the source text and between 1% and 3% in the target texts). Three human evaluators (interpreter trainers) from two universities were provided with the transcription of the interpreter's performance, the source text and the reference text. The same criteria used in the prompting were also used as a guideline and provided for human annotators beforehand. The evaluators were asked to evaluate the performance based on the given criteria and asked to give scores out of 10 for each criterion specified in 2.3.

Given the limited number of evaluators and the small sample size, the application of statistical models such as the Intraclass Correlation Coefficient (ICC), Pear-

son's correlation coefficient, and Fleiss Kappa would be inappropriate and result in a lack of generalizability. Given the limited number of evaluators and evaluation criteria, we think that, at the first stage, a qualitative analysis of the results based on scoring may be a more appropriate approach to gain insights into the performance of both human and LLM-based evaluation methods. Future research with a larger sample size and a broader range of evaluation criteria could benefit from using these statistical models to obtain more robust and comprehensive results. Though the pilot study employed a score-based design for the evaluation as well as a comparison with human judgement, the results section touches upon a rough analysis of the remarks. Perhaps, a detailed error identification analysis of the models can yield a better understanding. Further research can compare the detailed critics and points in the generated feedback and those of human evaluators by thematic classifications of the remarks.

3.2 Results

An overall overview of the remarks provided in the two outputs shows that GPT-4 produces more accurate feedback with more examples for both evaluations compared to GPT-3.5-based feedback (see the appendix). Scoring-wise, GPT-4 surpasses GPT-3.5 in its alignment with human judgment. Table 1 showcases the scores assigned by human evaluators and those produced by the two models from InterpreTutor.

Table 1. The scorings of each evaluator

	Criteria	Evaluator 1	Evaluator 2	Evaluator 3	InterpreTutor GPT-3.5	InterpreTutor GPT-4
Interpreter I	Accuracy, Fidelity & Completeness	7	6	5	6	6
	Cohesion and Coherence	8	7	6	2	6
	Terminology	6	7	6	5	7
	Disfluency Markers	6	6	7	5	6
Interpreter II	Accuracy, Fidelity & Completeness	4	5	2	2	6
	Cohesion and Coherence	5	3	3	2	6
	Terminology	3	5	4	4	7
	Disfluency Markers	5	8	8	6	8

GPT-3.5 Results

- a) **Accuracy, Fidelity & Completeness:** Both the GPT 3.5 model's score and the GPT-4's score (6) are in the same range as the human evaluator scores (5, 6, and 7) for the performance of Interpreter I. This indicates that the LLM evaluation is closely aligned with the human evaluators' assessment of accuracy. However, a rough human evaluation of the outputs indicates the output of GPT-3.5 does not fully catalog all inaccuracies and omissions but gives a couple of accurate exam-

ples.¹⁴ The less number of examples of inaccuracies make the report less persuasive and missing. Repetitive evaluation sessions with GPT-3.5 keeps the score in the same range, but the examples differ each time, though they are correct takeaways. On the other hand, GPT-4-based output constituted more detailed feedback with correct points, making the evaluation fairer given such an average score. For the second interpreter, It's noticeable that the GPT-4 version consistently rated higher than the GPT-3.5 version and gave higher scores than human evaluators. This could suggest an overestimation by the LLM or a difference in the standards used for evaluation between the LLM and human evaluators.

- b) Cohesion and coherence:** For the Interpreter I, the GPT-3.5 score (2) is significantly lower than the human evaluator's scores while the GPT-4 score matches well with the human judgment. However, for the interpreter II, neither the human evaluators nor the LLMs seem to have a common ground. The GPT-3.5 score remains quite lower than the human judgment. On the contrary, the GPT-4 score keeps relatively high for this criterion (6).
- c) Terminology:** For the interpreter I, the GPT-3.5 score (5) is slightly lower than the human evaluator scores (6 and 7), while GPT-4 has given a closer score (7). This result indicates that, for the interpreter I, the GPT-4 evaluation is somewhat aligned with the human evaluators' assessment of terminology consistency, but it may not be as accurate as in the accuracy criterion. For the second interpreter, however, GPT-4 appears to evaluate the use of terminology more favourably than either the human evaluators or GPT-3.5 and GPT-3.5 score is more aligned with those of human evaluators.
- d) Disfluency Markers (filled pauses and repetitions only):** For the first interpreter performance, the GPT-3.5 score (5) falls within the range of the human evaluator scores (6 and 7) while the GPT-4-based evaluation matches the scores of two human evaluators (6). For the second performance, human evaluators have some divergence in their evaluations. The GPT-4 evaluation is reasonably more aligned with the two human evaluators' assessment of disfluency markers (8). The written feedback-wise, it produced more examples of filled pause annotations, especially for the Interpreter II, which makes the score look more reliable. However, the GPT-3.5-based evaluation indicated some hallucinations. This conclusion is based on the output (Interpreter I) that indicates that “there were several instances of filled pauses”. While remarks on filled pauses held true, there was only one case of repetition in the interpretation.

4 Conclusion and Future Developments

In this preliminary work, we aimed to investigate the capabilities of a pre-trained LLM, the GPT-4 and GPT-3.5 in the assessment and evaluation of interpreting performance on a textual level based on human criteria. In doing so, we designed a prompt based on the assessment criteria views on the relevant literature and proposed

¹⁴ The examples can be seen in the Appendix 1.

an all-in-one LLM-powered application “InterpreTutor” which streamlines the process by using speech-to-text technology (OpenAI Whisper) and speech translation (OpenAI Whisper Translate). Whisper ASR, thanks to its state-of-the-art architecture and extensive training data, is successful at providing robust and clean ASR outputs¹⁵ which makes it easy to process by the GPT API. In our small-scale pilot test, we designed a prompt and made two interpreters’ performances evaluated by the LLM in InterpreTutor and three human evaluators.¹⁶ We observed that for the first interpreting performance, the GPT-4 model performed closely to human evaluators in terms of scoring compared to GPT-3.5. However, for the second interpreter, GPT-4 tended to assign higher scores than GPT-3.5 and sometimes even higher than human evaluators. GPT-3.5 generally aligned more closely with the lower range of human evaluator scores. Though the first test has shown a similarity in scoring between human evaluators and LLMs (particularly GPT-4), the second test provided an evident discrepancy not only between different models but also between models and human judgement, which gives no room for a definite conclusion. Despite the good alignment between human judgement and the InterpreTutor, the usability of such a tool for scoring is still questionable because of the inconsistency and the reliability of the scores. Nevertheless, InterpreTutor with GPT-4 performed well at analysing and commenting more persuasively and elaboratively on the audio inputs transcribed by the Whisper. Detailed feedback generation and accurate error detection of the models hold promise. Further optimization in prompting and fine-tuning of the model can generate useful output. LLMs, at this stage, have the potential to offer in-depth, partly constructive evaluation reports on the content accuracy with their sophisticated natural language understanding capabilities. Applications like InterpreTutor can be used by trainees in their deliberate-practice sessions when they want to independently assess themselves in some aspects.

To advance the understanding and application of LLMs in the assessment of translation and interpretation, the following areas can be recommended for future research. First, more empirical research with fine-tuned LLM models using larger datasets is recommended, as this could potentially lead to improved performance and enhanced evaluation capabilities. Second, this study focused on its feasibility in interpreter assessment and evaluation, while a similar study can be adapted for written translation assessment for scoring, grading or other pedagogical purposes. Even so, with the combination of LLMs’ good reasoning capacities and real-time speech-to-text/text-to-speech technologies, similar applications can be developed for the automatic evaluation of standardized speaking exams or similar simulations. This could be done by instructing the LLM to conduct a semantic, syntactic analysis and analyse the argumentation in the speech. However, a lot of development and testing would surely be

¹⁵ The study does not employ detailed benchmarking for ASR outputs except Word-error-rate calculation (which was shown to be very low). It should be acknowledged that ASR errors can also affect the output. Though most of the points made are correct, the model may occasionally misinterpret the text as a result of errors introduced by the ASR. Any lexical error in the ASR results can make the tool detect it as a major error.

¹⁶ This result must be interpreted in a nuanced manner, as it is subject to certain limitations.

required. Third, the current study focused on the Turkish-English language pair, and it would be beneficial to investigate the performance of LLMs in assessing translations and interpretations in other, particularly high-resource, languages. Fourth, the present study utilized a low-temperature setting for the GPT-3.5 due to the 4,000-token limitation imposed by the API at the time of writing, whereas GPT-4 has the capability of using more than 4k tokens which allows us to make it generate logical and longer evaluation reports. Future research should experiment with various temperature levels, different base models, few-shot prompting methods and larger token capacities to determine the optimal configuration for achieving the best results. Fifth, our study did not reveal the superiority of GPT-4 over GPT-3.5. Still, we observe that the GPT-4-based system for the workflow performs the best results in automatic assessment in the context of qualitative feedback generation. That's why, rather than just comparing numerical scores, analyzing the specific content of the feedback could offer a more nuanced understanding of how these two types of evaluators differ in their assessments and how Interpretutor can use the instructional capabilities of the pre-trained LLMs.

Certain restrictions and disadvantages are present since GPT-4 is slower and more costly. On top of that, ethical considerations on trustworthy usage, privacy, etc. do apply for Interpretutor. Task-specific prompting keeps the model away from hallucinating and hazardous content. However, it is also true that misleading information or misevaluations can rarely happen in the feedback output, particularly when using low-resource languages for source and target languages. This is why using a gold standard as a reference helps the model to report well-reasoned points in the evaluation. Language variations pose a challenge to the robustness of the transcribed content. Experimenting with different language combinations and even with more and more large-scale language models would undoubtedly unveil more about the practicality of such secondary training-purposed tools like Interpretutor.

References

1. Wang, L., et al.: Document-Level Machine Translation with Large Language Models. arXiv preprint arXiv:2304.02210 (2023).
2. Vaswani, A., et al.: Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).
3. Brown, T., et al.: Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33, pp. 1877–1901 (2020).
4. Eloundou, T., Manning, S., Mishkin, P., & Rock, D.: GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. arXiv preprint arXiv:2303.10130 (2023).
5. Elsen-Rooney, M.: NYC education department blocks ChatGPT on school devices, and networks. *Chalkbeat New York* January 3, 2023. <https://ny.chalkbeat.org/2023/1/3/23537987/nyc-schools-ban-chatgpt-writing-artificial-intelligence> (2023).

6. Kuzman, T., Mozetic, I., & Ljubešić, N.: ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification. *ArXiv abs/2303.03953* (2023).
7. Luo, Z., Xie, Q., & Ananiadou, S.: ChatGPT as a Factual Inconsistency Evaluator for Abstractive Text Summarization. *arXiv preprint arXiv:2303.15621* (2023).
8. Kocmi, T., & Federmann, C.: Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520* (2023).
9. Vilar, D., et al.: Prompting PaLM for Translation: Assessing Strategies and Performance. *arXiv preprint arXiv:2211.09102* (2022).
10. HENDY, A., et al.: How good are gpt models at machine translation? A comprehensive evaluation. *arXiv preprint arXiv:2302.09210* (2023).
11. Jiao, W., et al.: Is ChatGPT a good translator? A preliminary study. *arXiv preprint arXiv:2301.08745* (2023).
12. Peng, K., et al.: Towards Making the Most of ChatGPT for Machine Translation. *arXiv preprint arXiv:2303.13780* (2023).
13. Goyal, T., Li, J.J., & Durrett, G.: News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356* (2022).
14. Wang, J., et al.: Is chatgpt a good NLG evaluator? A preliminary study. *arXiv preprint arXiv:2303.04048* (2023).
15. Hua, M., & Raley, R.: Playing With Unicorns: AI Dungeon and Citizen NLP. *DHQ: Digital Humanities Quarterly* 14(4) (2020).
16. Gao, C.A., et al.: Comparing Scientific Abstracts Generated by CHATGPT to Original Abstracts Using an Artificial Intelligence Output Detector, Plagiarism Detector, and Blinded Human Reviewers. *bioRxiv* 2022.12.23.521610. (2022).
17. Han, C.: Interpreting testing and assessment: A state-of-the-art review. *Language Testing* 39(1), 30–55 (2021a).
18. Barik, H.C.: A Description of Various Types of Omissions, Additions and Errors of Translation Encountered in Simultaneous Interpretation. *Meta: Journal Des Traducteurs* 16(4), 199–210 (1971).
19. Setton, R., & Motta, M.: Syntacrobatics: Quality and reformulation in simultaneous-with-text. *Interpreting* 9(2), 199–230 (2007).
20. Schjoldager, A.: Assessment of Simultaneous Interpreting. *Teaching Translation and Interpreting* 3, p. 187– (1996).
21. Lee, J.: Rating Scales for Interpreting Performance Assessment. *The Interpreter and Translator Trainer* 2(2), 165–184 (2008).
22. Han, C.: Analytic rubric scoring versus comparative judgment: A comparison of two approaches to assessing spoken-language interpreting. *Meta* 66(2), 337–361 (2021b).
23. Lu, X., & Han, C.: Automatic assessment of spoken-language interpreting based on machine-translation evaluation metrics: A multi-scenario exploratory study. *Interpreting* 25(1), 109–143 (2022).
24. Yu, W., & Van Heuven, V.J.: Predicting judged fluency of consecutive interpreting from acoustic measures: Potential for automatic assessment and pedagogic implications. *Interpreting* 19(1), 47–68 (2017).
25. Han, C., et al.: Modeling the relationship between utterance fluency and raters' perceived fluency of consecutive interpreting. *Interpreting* 22(2), 211–237 (2020).
26. Ouyang, L., Lv, Q., & Liang, J.: Coh-Metrix model-based automatic assessment of interpreting quality. In: Chen, J., Han, C. (eds) *Testing and assessment of interpreting: Recent developments in China*, pp. 179–200. Springer, Singapore (2021).

27. Stewart, C., et al.: Automatic estimation of simultaneous interpreter performance. arXiv preprint arXiv:1805.04016 (2018).
28. Han, C., & Lu, X.: Can automated machine translation evaluation metrics be used to assess students' interpretation in the language learning classroom? *Computer Assisted Language Learning*, 1–24. (2021).
29. Wu, Z.: Chasing the unicorn? The feasibility of automatic assessment of interpreting fluency. In: Chen, J., Han, C. (eds) *Testing and assessment of interpreting: Recent developments in China*, pp. 143–158. Springer, Singapore (2021).
30. Le, N.-T., Lecouteux, B., & Besacier, L.: Automatic quality assessment for speech translation using joint ASR and MT features. arXiv preprint arXiv:1609.06049 (2016).
31. Stewart, C., et al.: Automatic estimation of simultaneous interpreter performance. arXiv preprint arXiv:1805.04016 (2018).
32. Li, X., et al.: Neural-based automatic scoring model for Chinese-English interpretation with a multi-indicator assessment. *Connection Science* 34(1),1638–1653 (2022).
33. Radford, A., et al.: Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356 (2022).
34. Heikinheimo, H.: Analyzing Open AI's Whisper ASR Accuracy: Word Error Rates Across Languages and Model Sizes. *Speechly*, 4 Apr. 2023, www.speechly.com/blog/analyzing-open-ais-whisper-asr-models-word-error-rates-across-languages, last accessed 2023/04/05
35. Riccardi, A.: Evaluation in interpretation. In: Hung, E., (eds.) *Teaching translation and interpreting* 4, pp. 115–126. John Benjamins Publishing Company, Netherlands (2002).
36. Dawrant, A., & Setton, R.: *Conference interpreting: A trainer's guide*. John Benjamins Publishing Company, Netherlands (2016).

APPENDIX I

GPT-4 Evaluation (Interpreter I)

Accuracy, Fidelity & Completeness:

a) *In general, the interpreter's translation conveys the meaning of the source text, but there are some misinterpretations and inaccuracies. Another example is "Avrupa ve Güney Afrika'da... Avrupa Birliği... Avrupa'da ve Güney Amerika'da en çok kazanılmıştır." which is a confusing and inaccurate representation of the source text. The source text states that only teams from Europe and South America have won the tournament, but the interpreter's translation mentions Europe and South Africa, and also brings up the European Union, which is not relevant to the source text.*

b) *There are several omissions and deviations in the interpreter's translation. Some examples include:*

- *The interpreter did not mention the fact that the FIFA World Cup is the most widely viewed sporting event.*
- *The interpreter omitted the specific number of titles won by Argentina, France, and Uruguay, as well as the fact that England and Spain have won the tournament once each.*
- *The interpreter did not mention the specific years and achievements of the African Federation teams (Cameroon, Ghana, and Senegal) reaching the quarterfinals.*
- *The interpreter did not mention the Women's World Cup following the same format as the Men's Tournament, and incorrectly stated that England has won the Women's World Cup twice, while the source text states that Germany has won twice.*
- *The interpreter did not mention the FIFA Under-20s and Under-17s World Cups being held every two years and allowing the best youngster players and national teams to compete against each other by age group.*

Cohesion and coherence: *The interpreter's translation is generally coherent, but there are some issues with cohesion and fragmentation. For example, the interpreter mentioned "Güney Afrika'da..."*

instead of "Güney Amerika'da...". There are also some filled pauses like "..." which disrupt the flow of the interpretation.

Terminology: In general, the interpreter used correct terminology, but there were some inconsistencies and errors. For example, the interpreter mentioned "Güney Afrika'da" instead of "Güney Amerika'da" and "Amerika Birliği" instead of "Avrupa Birliği." Additionally, the interpreter referred to the Women's World Cup as "Kadınların şampiyonası," which is not the exact term for the event.

Disfluency Markers: There are several instances of filled pauses in the interpreter's translation, such as "Avrupa ve Güney Afrika'da...", "Bu prestijli turnuva 3 milyar kişi tarafından izlendiği zaman...", and "Bir çok skandala imza atılsa da bu turnuvalar yine futbol sevenler tarafından...". These pauses disrupt the flow of the translation and make it less coherent. Additionally, there are repetitions like "Amerika kütasından katılan ülkeler olarak"

"32 takım... 32 takım yarışmaktadır." "Avrupa ve Güney Afrika'da... Avrupa Birliği... Avrupa'da ve Güney Amerika'da en çok kazanılmıştır." These disfluency markers indicate that the interpreter had difficulties in maintaining a smooth flow during the interpretation.

GPT-4 Evaluation (Interpreter II)

Accuracy, Fidelity & Completeness:

a) In general, the target text conveys the meaning of the source text, but there are some misinterpretations and inaccuracies. Some examples of major inaccuracies are:

- - The interpreter incorrectly states that the first tournament took place in 1910, while the source text says it was in 1930.
- - The interpreter mentions that Argentina has 3 championships, while the source text states that Argentina has 2 championships.
- - The interpreter incorrectly states that the USA's best result was in 2013, while the source text says it was in 1930.
- - The interpreter mentions Kemerun as an African country, which is not mentioned in the source text and is not a real country. The correct countries mentioned in the source text are Cameroon, Ghana, and Senegal.

b) There are some deviations and omissions in the interpreter's translation. Some examples are:

- - The interpreter omits the mention of Italy and Germany having four championships each.
- - The interpreter omits the mention of France and Uruguay having two championships each.
- - The interpreter omits the mention of England and Spain having one championship each.
- - The interpreter omits the mention of South Korea's fourth-place finish in 2002 being the best performance by an Asian nation.
- - The interpreter omits the mention of the Women's World Cup following the same format as the Men's Tournament.
- - The interpreter omits the mention of the FIFA Under-20s and Under-17s World Cups being held every two years.

Overall Rating Score: 6/10

Cohesion and coherence: Overall, the target text is logically connected and coherent, but there are some fragmentation issues and minor errors. For example, the interpreter mentioned the first World Cup took place in 1910, which is incorrect; it was in 1930. Also, the interpreter mentioned Argentina having three titles, while the correct number is two. Despite these errors, the target text still makes sense as a whole and maintains a narrative in Turkish.

Score: 7/10

Terminology: In general, the interpreter used correct and consistent terminology. However, there are some major errors:

- - The interpreter mentioned the first World Cup took place in 1910 instead of 1930.
- - The interpreter mentioned Argentina having three titles instead of two.
- - The interpreter mentioned the USA's best result was in 2013 instead of 1930.
- - The interpreter mentioned Kemerun instead of Cameroon.
- - The interpreter mentioned the best result for African countries was in 2006 instead of the correct information about Cameroon, Ghana, and Senegal reaching the quarterfinals.

Score: 6/10

Disfluency Markers: There are several occurrences of repetitions and filled pauses in the interpreter's performance. Examples include:

- - "Ve ilmiden fazla da türü var." (repetition of "ve")
- - "Sonrasında Almanya ve İtalya takip ediyor onu 4 şampiyonlukla." (repetition of "sonrasında")

- - "Sonrasında Kemerungan'a ve Senegal gibi Afrika ülkelerinde de en iyisi 2006 yılında gerçekleşmişti." (repetition of "sonrasında")
 - - "Sonrasında Avustralya için de durum bu şekilde devam ediyor." (repetition of "sonrasında")
 - - "Tabii ki erkek futbol turnuvası olan FIFA'daki bu erkek futbol turnuvasına çok fazla ilgi var." (repetition of "erkek futbol turnuvası")
- Score: 8/10

Google Translate vs. ChatGPT: Can non-language professionals trust them for specialized translation?*

Lucía Sanz-Valdivieso¹[0000-0001-5772-8041] and Belén López-Arroyo²[0000-0002-9171-1910]

^{1,2} University of Valladolid, Spain. P. Campus s/n, 47011, Valladolid (Spain)
lucia.sanz.valdivieso@uva.es

Abstract. Experts and professionals in specialized fields often need writing tools to communicate in English as a means to disseminate their knowledge or enter the international market. There are different tools to accomplish this and most of them are, lately, Machine Translation systems (MT) based on Neural Machine Translation (NMT), an approach using artificial neural networks to translate with outstanding fluency. Free and open systems such as Google Translate or, more recently, ChatGPT used as a translator, have popularized NMT to a multitude of users. However, there are experts and professionals who, due to their lack of command of English, often fail in their communication tasks by accepting NMT system's output as correct. This paper examines these systems' performance when translating terminology of the discourse in wine and olive oil tasting notes, specifically from Spanish into English. This domain may serve to represent less-studied specialized languages where general language words and terms become closely intertwined. The aim is to determine whether these systems can translate terminology accurately within the domain, and, if so, whether the GPT-3.5 model outperforms Google Translate. Results will help identify or discard possible language solutions for users who need to obtain texts in specialized English with professional and internationalization purposes, but who do not have the linguistic or economic resources to ensure the quality of the English text. Results show that, although ChatGPT yields fewer terminological errors than Google Translate in terms of error severity and number of samples affected, professionals cannot rely solely on these tools just yet.

Keywords: Languages for Specific Purposes, Terminological accuracy, Translation Quality Assessment.

* The authors belong to the ACTRES (*Análisis Contrastivo y Traducción Especializada ES-EN*) research group. This paper has been written within the research project *Lenguajes naturales controlados, comunicación colaborativa y producción textual bilingüe en entornos 3.0* (PID2020-114064RB-I00), supported financially by the *Ministerio de Educación y Ciencia de España*, and the project *Writing Audit: Evaluación de la redacción técnica con entornos visuals* (PID-078) supported financially by the University of Valladolid. Lucía Sanz-Valdivieso develops her research under a fellowship granted by the *Ministerio de Educación* (FPU20-00293).

1 Introduction

Since the computer was invented, humans have been, rather illusorily, aiming at Fully Automatic High Quality Machine Translation (FAHQMT) [1]. However, there has not been a model closer to that aim than Neural Machine Translation (NMT), the indisputable state-of-the-art in the field of MT. Its main advancement regarding its predecessors lies on its computational approach and its immensurable potential only limited by computer power and memory [2]. The neural approach to Natural Language Processing (NLP), based on Artificial Neural Networks (ANNs) [3], allows NMT to account for the richness of language through the principle of semantic compositionality and vectorial representation. Thus, NMT systems “build the interpretation of each sentence by combining the individual interpretations of its component words” [4, pp. 141–142].

1.1 Relevant work

Indeed, the wake of new applications of this technology into chatbots, which can be used for translation, may become a threat for some professionals, including translators [5]. There are even voices claiming that NMT systems can produce translations of such a high quality that “might and should worry some translators ... [b]ecause it is close to FAHQMT” [6, p. 201]—so much so, that there have already been declarations of MT reaching human parity [7]. However, these systems are still far away from attaining FAHQMT in the majority of text types, language combinations, and when the source text is not written in a Controlled Language, although this situation is rapidly changing [6]. In fact, while NMT’s general quality is higher than other systems—in terms of fluency, accuracy, but not style—, this is not perceivable in all language pairs and it is negatively affected by sentence length [8; 9; 10]. Errors of any kind, especially critical errors, increase when translating online user-generated content, which is usually colloquial, ungrammatical, and contains emojis and other characters [11].

Still, numerous studies point towards NMT as the highest quality kind of MT in different pair combinations and using different assessment methods in high-to-moderate resource settings [1; 12]. While general NMT quality is indeed higher than Statistical MT [8; 9; 10; 13; 14; 15; 16; 17; 18], most errors tend to be lexical [19], even though NMT produces fewer word order and morphological errors [16]. NMT also outperforms Phrase-Based MT in technical translation quality in the business language, except in the categories of terminology and formatting tags [20; 21; 22; 23]. Post-editing effort is also lower in NMT systems’ output [24; 25]—the most frequent changes are related to word substitutions and word form [24], confirming too NMT’s relative terminological and lexical weakness [26]. This is especially relevant since many texts to be translated using MT belong to specialized domains, where terminology takes a central role [27]. Comparing two free open-source NMT systems, Google Translate and DeepL, when translating Spanish phraseological units both show a similar performance which is weakened when encountering low-frequency expressions [28]. Other studies confirm such results in Portuguese-French, where phraseology, calque and nonsense were the most frequent errors [29].

1.2 Motivation and focus of the study

The purpose of this paper is to test the terminological accuracy of the latest NMT systems in a very specific text type within a specialized discourse: olive oil and wine tasting notes written in Spanish. Tasting notes are usually short texts describing a product's organoleptic attributes, composed of relatively long sentences and full of lexical and terminological richness [30; 31; 32; 33]. The wider study in which this experiment develops addresses a very common type of user, i.e., professional or technical experts who, in spite of not being able to produce specialized texts in English by themselves, do need to obtain such texts. These users need texts in English as the international *lingua franca* for a diversity of purposes, ranging from marketing, to labelling, to touristic promotion and education. These factors ultimately determine the international economic performance of sectors as important as the wine and olive oil industries in Spain, in this case. Nevertheless, the ongoing project aims at extrapolating results to other specialized fields where speakers belong to small-medium organizations and need to obtain English texts but do not have the ability to compose such texts on their own nor the means to adopt quality language services.

The relevance of the kind of expert described above is in their lack of ability to identify an inadequate translation. This would not pose a problem if these users' aim when using MT was gisting-related [3]—but these users know the content of the source text and are translating into a language they do not fully command. An added issue to this profile is the lack of economic means that most multi-national companies can invest in high-quality Language for Specific Purposes (LSP) translations to promote their internationalization. The reality is that these experts cannot possibly post-edit a faulty translation in the way a professional translator would. Rather, they will usually just copy and paste the MT system's output, or directly integrate a Google Translate plugin in their website to be able to offer its English version in some way, even if flawed (Fig. 1 below).



Fig. 1. Examples of resources currently used by these users to translate content into English.

This is just an instance showing how experts in this kind of small specialized domain take NMT as FAHQMT, even though translation professionals and English-speaking members of the discourse community would identify possible errors in the text [34]. While errors most frequently result in unnatural expressions unrecognizable for the target discourse community, they may also reach the extent of impeding successful

communication. Unfortunately, this may have serious consequences for companies individually but also for the sector as a whole [35; 36].

In this sense, few works have examined NMT in specialized contexts, with the exception of some works on political [37; 38] and biomedical discourse [39], which do not focus on terminology from a user perspective. However, those are very different from the LSP and text genre we are concerned with, which belongs to the tasting domain, a highly subjective field whose most representative text type is the tasting note. Tasting notes (TNs) may indeed be viewed as suitable candidates for FAHQMT given their short length and the frequency of agricultural and plant-related language in general NMT systems' training data. Presumably, it would be easier to translate expressions such as “*notas de plátano*” [banana notes] or “*hierba recién cortada en nariz*” [freshly cut grass on the nose] than terms from other more innovative and technology-related fields, such as “*apertura de la barrera hematoencefálica mediante ultrasonido focalizado*” [focused ultrasound blood-brain barrier disruption]. Hence, the tasting domain will serve to test NMT systems' terminological performance in LSPs which may not be extremely recent, technological or ubiquitous around the globe, but which certainly play a central role in countries' economies and will add to our understanding of how terminology is handled by NMT systems.

2 Methodology

2.1 Dataset

The dataset used consists of samples from olive oil and wine TNs Spanish corpora compiled for other related projects within the ACTRES research group [40]. The compilation followed pragmatic criteria: TNs were selected to ensure a representative sample of the language of expert members of the discourse community. TNs published by olive oil presses or wineries were taken from the websites registered to official and institutional sites such as *Aceites de Oliva de España* [41] and the *GOP Ribera de Duero* [42]. Samples were chosen randomly into the dataset to be translated using NMT systems. The number of texts selected from the corpus is determined by Biber's criterion of needing at least 20 samples of 2,000-5,000 words for a dataset to be representative of the register under study [43, p. 261]. This experiment uses 25 samples from each corpus, amounting to a relatively small but specialized dataset of 50 samples (5,122 tokens total): the olive oil TNs sample contains 2,577 tokens (699 types, 737 lemmas), and, the wine TNs sample, 2,545 tokens (789 types, 817 lemmas).

The selection of MT system(s) to be tested is determined by the purpose of each work [44], where mathematical linguists tend to train specific systems, while other more purely linguistic projects focus on commercial systems [18]. This study makes use of Google Translate NMT (GNMT) system [45] and ChatGPT (CGPT-3.5) [46] accessed in April 2023—these systems are free, open and popular, so they could be expected to throw similar results as when real life users use them to translate their Spanish TNs into English. In the case of CGPT-3.5, the prompt “*Traduce de español a*

inglés” [Translate from Spanish into English] was used for it to act as a translator without finetuning the results through a more accurate prompt—a path currently under study in the wider project where this experiment develops.

2.2 Methods

To test the terminological accuracy of current NMT systems at translating TNs, this paper aims at analyzing translations performed by free, open and popular systems, as part of the project which considers complementary means of human Translation Quality Assessment (TQA) as well as automated metrics. For this purpose, we used our familiarity with the LSP of olive oil and wine tasting in English and our background in translation and linguistics to perform the human evaluation of the target texts from both NMT systems. Hence, there were two annotators in this experiment, where inter-annotator agreement was calculated through Cohen’s Kappa with a result of $K= 0.7242$, indicating substantial agreement.

This TQA was performed through the Multidimensional Quality Metrics (MQM) framework, developed to provide a comprehensive and standardized quality assessment model [36]. It comprises a set of 182 issue types hierarchically organized into dimensions; not all of the types are to be covered in the assessment of a translation—rather, they are to be used to ensure that said translation “meets specifications” [47, p. 119]. In other words, the MQM proposes a functionalist framework where the translation’s purpose in context plays a central role in how its quality should be assessed [48]. While the broader project on which this work-in-progress paper is based covers more MQM aspects, only a limited set of issues are reported here.

The terminology dimension of the MQM “relate[s] to the use of domain- or organization-specific terminology” and is made up of three possible kinds of issues: inconsistency with termbase, with domain, or inconsistent use of terminology along the text [49]. Our analysis focuses on terminological inconsistency with domain, since there is not a specific termbase that standardizes tasting terminology. Hence, a term was flagged in a translation when it “is used contrary to general domain expectations” [49]. Nevertheless, a forthcoming terminological and phraseological glossary we developed for related projects was consulted for guidance when necessary. In addition and for the sake of accounting for terminological accuracy comprehensively, the issue “mistranslation” within the “accuracy” dimension (i.e., “does not accurately represent the source content” [49]) was also considered when terminology was affected. Untranslated expressions were noted as well.

To quantify the extent of the issues detected through the selected parameters, the MQM provides four severity levels: critical errors, where a translation is unfit for its purpose, involving legal, safety or usability consequences; major errors, which “make the intended meaning of the text unclear ...[and] the user cannot recover the meaning” [47, p. 120]; minor errors, without an impact on usability; and null level (changes that are not errors) [47]. Since tasting notes hardly ever entail critical danger, level 1 is excluded from the analysis, as well as level 4, since we are concerned with errors *per se*.

3 Results and discussion

After following the MQM to examine the 50 samples and their translations, the analysis revealed that samples translated by CGPT-3.5 yield 21.57% fewer errors, mistranslations and untranslated elements of the tasting LSP than GNMT:

Table 1. Terminological issues detected through the selected MQM parameters

MT system	Terminological inconsistency with domain		Mistranslated expressions		Untranslated expressions	Total
	Major	Minor	Major	Minor		
GNMT	23	50	3	15	11	102
CGPT-3.5	14	45	1	8	12	80

Results seem to show CGPT-3.5 used as a translator outperforms GNMT in terms of general terminological accuracy when working with TNs from Spanish into English. Most errors belong to the terminological issue type, i.e., translations that could work outside the specific domain in question. In many cases, both systems overlooked domain-specific terminology and used other general language equivalents instead:

1. Source text: ... *un vino de capa alta, de gran brillantez*, ...
GNMT: ... *a wine with a high robe, of great *brilliance*, ...
Source text: ... *el aceite presenta un aspecto brillante*.
CGPT-3.5: ... *the oil has a *shiny appearance*.

Other common terminological errors found in the translations include “*capa*,” translated as “*layer*” and not as “*robe*” in most of the cases by both systems; “*entrada*” as “*entrance*” and not “*entry*”; “*paso*” and “*recorrido*” often translated as “*step*”, “*passage*” and “*journey*” instead of “*mid palate*”; “*recuerdos*” as “*memories*” or “*reminders*” and not as “*hints*”, “*notes*”, or even “*reminiscences*”; or the color descriptor “*teja*” translated as “*tile*” instead of “*brick*”; which CGPT-3.5 used interchangeably.

Tasting verbs were also not correctly translated, such as “*ofrece*” or “*regalar*” being translated as “*give*” instead of “*offer*”, or progressive expressions such as “*apreciándose*” being literally translated as “*appreciating*”. Other mistranslations include “*aceituna de pre-envero*” translated as “*pre-veraison olive*” (GNMT) or “*olive in pre-winter*” (CGPT-3.5) and not “*green olive*”. Untranslated expressions were recurrently “*alozza*” [green almond] and “*bodega*” [cellar, winery, vineyard].

However, CGPT-3.5 was slightly more accurate than GNMT:

2. Source text: *El picor es ligero pero se nota*.
GNMT: *The *itch is slight but it is noticeable*.
CGPT-3.5: *The pungency is light but noticeable*.

Other instances where only CGPT-3.5 was able to find the correct tasting term include simple agrarian terms such as “*tomatera*”, translated as “*tomato plant*” by CGPT-3.5 but as “*tomato*” by GNMT, which also output “*nariz voluminosa*” as “*bulky nose*”; or “*zumos cordobés* [from Córdoba, Spain]” as “*Cordovan juice*”. Similarly, CGPT-3.5 was able to correctly translate the verbs “*finalizar, terminar*” as “*finish*” and not “*end*”.

Yet, particularly note-worthy are GNMT’s incorrect translations of “*vista*” [appearance] as “*sight*” and “*view*”, and, “*nota de cata*” [tasting note], as “*Cata’s note*”.

Not only did CGPT-3.5 produce translations with fewer errors as a whole—it also output more texts free of any terminological errors, with a 38.89% difference from GNMT texts:

Table 2. Samples containing no, only minor, only major, and both minor and major errors.

MT system	Number of samples				
	Error severity levels	None	Minor	Major	Both
GNMT		7	21	6	16
CGPT-3.5		18	17	5	10

Still, even though 36% of the samples translated by CGPT-3.5 were error-free in contrast to the 14% of the texts by GNMT, this does not mean that they are ready to be regarded as acceptable. It is worth taking a closer look at one of the translations where no terminological issues were detected by GNMT (example 3) and CGPT-3.5 (example 4):

3. *Very greenish yellow color. The nose is intense, complex, fresh grass, tomato plant, artichoke hints and a touch of dried fruit. On the palate it is dense, fruity, of good intensity, spicy, clean, with a vegetal touch, well-balanced sweet sensations and good length.*

4. *Wine with an intense cherry red color and soft violet notes, which indicates its aging with bright terracotta and amber edges. A nose of great subtlety with aromas of ripe fruit well combined with the aging in wood, leading us to special aromas (vanilla, cinnamon) to the stimulating aroma of coffee or toasted notes. Very meaty on the palate, with a long finish and balanced acidity.*

These are accurate translations terminologically speaking, but both the fluency and some grammatical structures are questionable to different extents. For instances, in example 3, the main verb “is” is used to embed a series of phrases into an enumeration that results rather unnatural in English; while, in example 4, the first two sentences become so long—which is completely natural in Spanish—that even grammatical mistakes can be spotted (“lead *from X to Y*”, and not “lead *to X to Y*”). In this sense, there is ample room for improvement in most terminologically accurate translations, as well as in those with errors, in terms of fluency. In other words, even though NMT’s output sounds surprisingly more natural than previous systems’, these outputs prove how there are still robotic transfers of grammatical patterns into the target language that prevent them from being acceptable within the target discourse community. Besides, issues such as consistency are worth a more detailed analysis, since most errors were frequently, but not always present in both of the systems’ output.

4 Conclusion

This work-in-progress paper has focused on how the latest and most popular free and open MT systems treat terminology within the specialized field of tasting, which differs from other more pervasive, technological and more objective domains. This

rendered a sample of olive oil and wine tasting notes a rich and interesting ground for research in this regard—many human activities and sectors use specialized languages which are often too small, subjective or intermingled with general language for MT systems to be able to correctly translate them, even in two of the largest languages in the world. Our MQM-based, terminology-focused analysis has proved the enormous potential of these systems while revealing terminology is not the greatest strength of neural systems. Most importantly, results show a better performance by CGPT-3.5 used as a translator than GNMT terminology-wise.

Still, none of these systems outputs texts that are acceptable for the user who does not have the training, linguistic or economic means necessary. The wider project in which this work in progress develops overcomes some of the limitations of the present paper, and is currently looking at other aspects of the MQM, as well as applying automatic TQA metrics in order to obtain a complete picture of the behavior of these systems. In any case, there is a long way ahead in order to develop tools which can help this kind of user achieve their goal, which may range from finetuned MT systems to complementary tools such as terminological aids that may help them obtain their TNs in English and so promote and internationalize their products, businesses, and cultural assets.

References

1. Koehn, P., Knowles, R.: Six challenges for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation, pp. 28–39. Association for Computational Linguistics, Vancouver, Canada (2017).
2. Forcada, M. L.: Making sense of neural machine translation. *Translation Spaces* 6(2), 291–309 (2017).
3. Koehn, P.: *Neural Machine Translation*. Cambridge University Press, Cambridge (2020).
4. Pérez-Ortiz, J. A., Forcada, M. L., Sánchez-Martínez, F.: How neural machine translation works. In Kenny D. (ed.) *Machine translation for everyone: Empowering users in the age of artificial intelligence*, pp. 141–164. Language Science Press, Berlin (2022).
5. Eloundou, T., Manning, S., Mishkin, P., Rock, D.: GPTs are GPTs: An early look at the labor market impact potential of Large Language Models. *ArXiv* (2023).
6. Schmitt, P. A.: Translation 4.0 – Evolution, Revolution, Innovation or Disruption? *Lebende Sprachen* 64(2), 193–229 (2019).
7. Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., Zhou, M.: Achieving Human Parity on Automatic Chinese to English News Translation. Microsoft research, *arXiv* (2018).
8. Álvarez-Vidal, S., Oliver, A., Badia, T.: What do post-editors correct? A fine-grained analysis of SMT and NMT errors. *Revista Tradumàtica. Tecnologies de la Traducció* 19, 131–147 (2021).
9. Bentivogli, L., Bisazza, A., Cettolo, M., Federico, M.: Neural versus Phrase-Based Machine Translation quality: A case study. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, pp. 257–267. EMNLP, Texas, USA (2016).

10. Toral, A., Sánchez-Cartagena, V. M.: A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pp. 1063–1073. Association for Computational Linguistics, Valencia, Spain (2017).
11. Al Sharou, K., Specia, L.: A taxonomy and study of critical errors in machine translation. In: Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, pp. 171–180. Belgium, (2022).
12. Haddow, B., Bawden, R., Miceli Barone, A. V., Helcl, J., Birch, A.: Survey of low-resource Machine Translation. *Computational Linguistics* 48(3), 673–732 (2022).
13. Barrault, L., Bojar, O., Costa-Jussà, M., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., Zampieri, M.: Findings of the 2019 Conference on Machine Translation (WMT19). In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 1–61. Association for Computational Linguistics, Florence, Italy (2019).
14. Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névóel, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., Zampieri, M.: Findings of the 2016 Conference on Machine Translation. In: Proceedings of the First Conference on Machine Translation, pp. 131–198. Association for Computational Linguistics, Berlin, Germany (2016).
15. Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Monz, C.: Findings of the 2018 Conference on Machine Translation (WMT18). In: Proceedings of the Third Conference on Machine Translation Shared Task Papers. Association for Computational Linguistics, Brussels (2018).
16. Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., Way, A.: Is Neural Machine Translation the new state of the art? *Prague Bulletin of Mathematical Linguistics* 108(1), 109–120. (2017).
17. Ramesh, A., Parthasarathy, V. B., Haque, R., Way, A.: Comparing Statistical and Neural Machine Translation performance on Hindi-to-Tamil and English-to-Tamil. *Digital* 1, 86–102 (2021).
18. Rivera-Trigueros, I.: Machine translation systems and quality assessment: a systematic review. *Lang Resources & Evaluation* 56, 593–619 (2022).
19. Bentivogli, L., Cettolo, M., Federico, M., Federmann, C.: Machine Translation Human Evaluation: an investigation of evaluation based on Post-Editing and its relation with Direct Assessment. In: Proceedings of the 15th International Conference on Spoken Language Translation, pp. 62–69. International Conference on Spoken Language Translation, Brussels (2018).
20. Beyer, A., Macketanz, V., Burchardt, A., Williams, P.: Can out-of-the-box NMT beat a domain-trained Moses on technical data. In: *EAMT 2017: The 20th Annual Conference of the European Association for Machine Translation*. Prague, Czech Republic (2017).
21. Klubička, F., Toral, A., Sánchez-Cartagena, V. M.: Fine-grained human evaluation of Neural versus Phrase-Based Machine Translation. *The Prague Bulletin of Mathematical Linguistics* 108, 121–132 (2017).
22. Klubička, F., Toral, A., Sánchez-Cartagena, V. M.: Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian. *Machine Translation* 32, 195–215 (2018).

23. Vintar, S.: Terminology translation accuracy in Phrase-Based versus Neural MT: An evaluation for the English-Slovene language pair. In: Du, J., et al. (eds) Proceedings of the LREC 2018 Workshop “MLP–MomenT”, pp. 34–37. Miyazaki, Japan (2018).
24. Koponen, M., Leena, S., Nikulin, M.: A product and process analysis of posteditor corrections on neural, statistical and rule-based machine translation output. *Machine Translation* 33, 61–90 (2019).
25. Ye, Y., Toral, A.: Fine-grained human evaluation of transformer and recurrent approaches to Neural Machine Translation for English-to-Chinese. In: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pp. 125–134. European Association for Machine Translation, Lisboa, Portugal (2020).
26. Saunders, D.: Domain adaptation and Multi-Domain adaptation for Neural Machine Translation: A survey. *Journal of Artificial Intelligence Research* 75, 351–424. 2022.
27. Calvi, M. V., Bordonaba Zabalza, C., Mapelli, G., Santos López, J.: *Las Lenguas de Especialidad en Español*. Carocci editore, Roma (2009).
28. Hidalgo-Ternero, C. M.: Google Translate vs. DeepL: analysing neural machine translation performance under the challenge of phraseological variation. In: Mogorrón Huerta, P. (ed.) *Multidisciplinary Analysis of the Phenomenon of Phraseological Variation in Translation and Interpreting*. MonTI Special Issue 6, 154–177 (2020).
29. Bacquelaire, F.: DeepL and Google Translate Translating Portuguese Multi-Word Units into French: Progress, Decline and remaining Challenges (2019-2023). In: Vetulani, Z., and Paroubek, P. (eds.) *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pp. 19–23, 2023.
30. López-Arroyo, B., Sanz-Valdivieso, L.: Tasting notes: A corpus-based study of olive oil and wine tasting discourse. *Iberica* 43, 205–234 (2022a).
31. López-Arroyo, B., Sanz-Valdivieso, L.: The phraseology of wine and olive oil tasting notes: a corpus based semantic analysis. *Terminology* 28(1), 37–64 (2022b).
32. Sanz-Valdivieso, L., López-Arroyo, B.: *Generador de Notas de Cata de Aceite de Oliva 1.0: Lingüística Aplicada a la Internacionalización del Aceite de Oliva*. Diputación de Jaén, Jaén, Spain (2022).
33. Sanz-Valdivieso, L., López-Arroyo, B.: On Describing Olive Oil Tasting Notes. *Fachsprache. Journal of Professional and Scientific Communication* 42(1–2), 27–45 (2020).
34. Swales, J. M.: *Genre Analysis. English in Academic and Research Settings*. Cambridge University Press, Cambridge, UK (1990).
35. Castilho, S., Doherty, S., Gaspari, F., Moorkens, J.: Approaches to human and machine translation quality assessment. In Moorkens, J. et al. (eds.) *Translation Quality Assessment, Machine Translation: Technologies and Applications 1*, pp. 9–38 (2018).
36. Lommel, A., Uszkoreit, H., Burchardt, A.: Multidimensional Quality Metrics (MQM): A framework for declaring and describing Translation Quality metrics. *Revista Tradumàtica. Tecnologies de la Traducció*, 12, 455–463 (2014).
37. Liu, S., Zhu, W.: An Analysis of the Evaluation of the Translation Quality of Neural Machine Translation Application Systems. *Applied Artificial Intelligence*, 37(1), 1505–1531 (2023).
38. Vardaro, J., Schaeffer, M., Hansen-Schirra, S.: Translation Quality and Effort Prediction in Professional Machine Translation Post-Editing. In: Proceedings of the Second MEMENTO workshop on Modelling Parameters of Cognitive Effort in Translation Production, pp. 7–8. European Association for Machine Translation, Dublin, Ireland (2019).
39. Wang, W., Meng, X., Yan, S., Tian, Y., Peng, W.: Huawei BabelTar NMT at WMT22 Biomedical Translation Task: How we further improve domain-specific NMT. In: Proceedings

- of the Seventh Conference on Machine Translation (WMT 2022), pp. 930–935. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (2022).
40. ACTRES Research Group Homepage, <https://actres.unileon.es/wp/>, last accessed 2023/06/13.
 41. Aceites de Oliva de España [Olive Oils from Spain] Homepage, <https://www.aceitesdeolivadeespana.com/>, last accessed 2023/06/13.
 42. GOP Ribera de Duero Homepage, <https://www.riberadelduero.es/>, last accessed 2023/06/13.
 43. Biber, D.: Methodological issues regarding corpus-based analysis of linguistic variation. *Literary and Linguistic Computing* 5(4): 257–269 (1990).
 44. Rossi, C., Carré, A.: How to choose a suitable neural machine translation solution: Evaluation of MT quality. In Kenny, D. (ed.) *Machine translation for everyone: Empowering users in the age of artificial intelligence*, pp. 51–79. Language Science Press, Berlin (2022).
 45. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: *Google’s Neural Machine Translation system: Bridging the gap between human and Machine Translation*. ArXiv (2016).
 46. OpenAI Report on GPT-4, <https://openai.com/research/gpt-4>, last accessed 2023/04/18.
 47. Lommel, A.: Metrics for Translation Quality Assessment: A case for standardising error typologies. In Moorkens, J., et al. (eds.) *Translation Quality Assessment, Machine Translation: Technologies and Applications vol. 1*, pp. 109–128. Springer, Switzerland (2018).
 48. Nord, C.: *Translating as a purposeful activity*. St. Jerome, Manchester (1997).
 49. Multidimensional Quality Metrics (MQM) Issue Types, <https://web.archive.org/web/20211216145234/http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html>, last accessed 2023/04/18.

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

Michael Farrell¹[0000-0002-7138-6639]

¹ IULM University, Milan, Italy
michael.farrell@iulm.it

Abstract. This preliminary study consisted of two experiments. The first aimed to gauge the translation quality obtained from the free-plan version of ChatGPT in comparison with the free versions of DeepL Translator and Google Translate through human evaluation, and the second consisted of using the free-plan version of ChatGPT as an automatic post-editor of raw output from the pay-for version of DeepL Translator (both monolingual and bilingual full machine translation post-editing). The experiments were limited to a single language pair (from English to Italian) and only one text genre (Wikipedia articles).

In the first experiment, DeepL Translator was judged to have performed best, Google Translate came second, and ChatGPT, last.

In the second experiment, the free-plan version of ChatGPT equalled average human translation (HT) levels of lexical variety in automatic monolingual machine translation post-editing (MTPE) and exceeded average HT lexical variety levels in automatic bilingual MTPE. However, only one MT marker was considered, and the results of the post-editing were not quality-assessed for other features of MTPE that distinguish it from HT. It would therefore be inadvisable to generalize these findings at present.

The author intends to carry out new translation experiments during the next academic year with ChatGPT Plus, instead of the free-plan version, both as an MT engine and as an automatic post-editor. The plan is to continue to evaluate the results manually and not automatically.

Keywords: machine translation post-editing, human machine translation output evaluation, DeepL Translator, Google Translate, ChatGPT, automatic post-editing

1 Introduction

Although ChatGPT has only been available to the public since the end of November last year, some evaluation studies have already been carried out on the chatbot's ability to translate between natural languages, including Turkish, Romanian, Chinese, English and German [5, 6 and 7]. However, to the best of the author's knowledge, all of these

have so far used automatic metrics for raw output quality evaluation. Moreover, the author is unaware of any attempts to use ChatGPT as an automatic post-editor of machine translation (MT) output from other MT engines.

This preliminary study consisted of two experiments. The first aimed to gauge the translation quality obtained from ChatGPT in comparison with DeepL Translator and Google Translate, and the second attempted to use ChatGPT as an automatic post-editor of raw output from the pay-for version of DeepL Translator, examining both monolingual and bilingual full MT post-editing (MTPE).

The results were analysed to assess how to proceed with a new series of further-reaching experiments.

2 Design and methods

2.1 First experiment

Seven post-graduate students of translation (IULM University, Milan) comparatively assessed the raw output from the free-plan version of ChatGPT, based on the GPT-3.5 architecture, and the free versions of DeepL Translator and Google Translate. Three short extracts from the biographies of heterogeneous celebrities (Yehoshua Bar-Hillel [313 words], G. H. MacDermott [342 words] and Michael Jackson [358 words]) were taken from the English language version of Wikipedia and machine-translated into Italian on 6 April 2023. The three outputs were then segmented in Raw Output Evaluator¹ [3] and presented to the students, who simply assessed the three translations as *best*, *second best* and *worst* on a segment-by-segment basis (ties were allowed). A score was then calculated by assigning three points for each segment regarded as *best*, two points for *second best* and one point for *worst*. This simple ranking technique was chosen both for its speed and because the students had not yet received any training on the use of analytic metrics.

Wikipedia articles were used since they are likely to be less challenging for a machine translation system than classic works of literature but more problematic than the boilerplate-style texts that are normally considered to lend themselves best to machine translation.

The simple prompt used to generate the translation in ChatGPT was “Please translate the following text into Italian”, followed by a line break and then the source text.

2.2 Second experiment

A short extract from an English-language Wikipedia entry on Slovakia was taken for the second experiment (262 words). This text was chosen since it contained the bigram *there are* four times. This was the first short extract that contained at least four examples of the chosen n-gram in the space of approximately 250 words that the author came across while searching randomly through Wikipedia. Again, a Wikipedia entry was chosen in order to pose a medium-level challenge.

¹ www.intelliwebsearch.com/raw-output-evaluator

The bigram *there are* had been identified in a previous experiment carried out between 2017 and 2018 as among the best MT markers, that is n-grams that were translated with a highly statistically significantly greater number of correct translation solutions in human translation (HT) than in MTPE [2]. In the 2017/18 experiment, the frequency with which *there are* was translated into Italian with *ci sono* was successfully used to distinguish MTPE from human translation.

The aim of the second experiment was therefore to see if ChatGPT was able to post-edit MT output automatically and reduce the lexical impoverishment that has been observed to ensue from human post-editing [2 and 8]. Lexical uniformity is not a positive feature in texts that need to be engaging and intellectually stimulating, such as in the fields of marketing, advertising, literature, journalism, education, entertainment, and creative writing in general.

First, the author checked if the free-plan version of ChatGPT could be prompted to produce raw output in which the chosen MT marker occurred with a frequency that was in keeping with its frequency in HT. The prompt used was “Please translate the text below into Italian, keeping in mind that lexical variety is required for a human-quality final text.” This was followed by a line break and then “Here is the text in original language:”. After that, there was another line break, followed by the English language version of the Slovakia text.

Next the author asked ChatGPT to do automatic bilingual post-editing of raw output obtained from DeepL Translator. This MT engine was chosen because, in a recent survey among professional translators [4], the MT system most used by the respondents who declared that they use MT at some point in their workflow turned out to be DeepL Translator (183 users). Its nearest rival Google Translate was only chosen by just over half that number (93 users). Moreover, the majority of DeepL Translator users surveyed stated that they use the pay-for version (102).

The prompt used was “Please post-edit the text below, which was machine-translated into Italian, keeping in mind that lexical variety is required for a human-quality final text.” This was followed by a line break and then “Here is the text in original language:”. After that, there was another line break, followed by the English language version of the Slovakia text. This was then followed by “Here is the text to post-edit:”, a line break and then the raw output from the pay-for version of DeepL Translator.

Lastly the author asked ChatGPT to do automatic monolingual MTPE. The prompt used was “Please post-edit the text below, which was machine-translated into Italian, keeping in mind that lexical variety is required for a human-quality final text.” This was followed by a line break and then the same raw output as used before.

To establish the normal average frequency of the most chosen translation solution in HT for the MT marker *there are*, the same text on Slovakia was translated into Italian by 18 post-graduate students of translation [1]. The most frequent translation solution, *ci sono*, occurred 50% of the time. This solution occurred in the DeepL machine translated text three times out of four (75%). So, a human post-editor would tend to be primed to use this solution with a higher than natural frequency.

3 Results

3.1 First experiment

DeepL Translator was assessed to have performed best with a total score of 279 points, Google Translate came second with 239 points, and ChatGPT, last with 186 points. This preliminary result cannot however be generalized since it concerns only one language pair (English to Italian) and one text genre.

3.2 Second experiment

In the first part of the second experiment, ChatGPT was prompted to produce raw output in which lexical variety was required for *a human-quality final text*. However, ChatGPT failed to deliver: the MT marker *there are* was translated with the bigram *ci sono* four times out of four (100%), which is twice the previously measured average natural HT frequency in this text (50%).

ChatGPT was then asked to do automatic bilingual post-editing of raw output obtained from the pay-for version of DeepL Translator. This time, ChatGPT left the bigram *ci sono* as the translation of the MT marker *there are* only once despite being primed by the raw output with three occurrences. In other words, in the case of this specific MT marker and this specific text, ChatGPT produced greater lexical variety than the students did on average.

Lastly, ChatGPT was asked to do automatic monolingual post-editing of the same raw output. The result was two occurrences of the bigram *ci sono*. Therefore, ChatGPT reached human parity as far as the chosen MT marker is concerned in this particular monolingual post-editing.

4 Conclusion

Wenxiang Jiao et al. [6] report that the ChatGPT Plus version, based on GPT-4 architecture, scores higher than the free-plan version in automatic MT raw output evaluation metrics. Consequently, the planned future experiments will be carried out using ChatGPT Plus, and not the free-plan version.

Another limitation of the first experiment was that the evaluators knew which MT engine had been used to produce the raw output they were evaluating. Although it is unlikely that they expressed biased opinions on the basis of this knowledge, the future experiments will be carried out blind.

The prompt used to ask ChatGPT to translate the text in the first experiment does not take advantage of ChatGPT's ability to emulate different styles [7]. Better results may have been achieved with a prompt like "Please translate the following text into Italian in the style of a Wikipedia entry" or by providing information about the source text and purpose of the translation.

Seven human evaluators is a small number, which will be increased in the planned future experiments to reduce subjective biases. However, the author will in any case be limited by the size of the class for all experiments, which is unlikely to be much in

excess of forty students. Another limitation that cannot be overcome is the language pair (English to Italian). Academic time constraints will also limit the length and number of texts that may be analysed and the complexity of the analysis metrics.

In the first part of the second experiment, ChatGPT was prompted to produce raw output with human-like lexical variety. However, it failed to do so, at least in the case of the test MT marker chosen. Again, the result may have been different if ChatGPT Plus had been used. The same experiment should also be repeated on more than one text.

The most remarkable results were seen when ChatGPT was asked to post-edit raw output from the pay-for version of DeepL Translator. In the case of the specific MT marker considered and with the particular text chosen, ChatGPT reached average human-level lexical variety in monolingual MTPE and exceeded it in bilingual MTPE.

It is a little unfair to tell ChatGPT to consider lexical variety and not give the same instruction to the human post-editors. In future experiments, it might be interesting to divide the human post-editors into two groups and ask half of them to bear lexical variety in mind.

The automatic post-editing output produced by ChatGPT also needs to be evaluated to see to what extent a further stage of human post-editing is required. Again, future experiments will be carried out with ChatGPT Plus.

Interestingly, the author has recently received an offer from a language service provider, based in Hong Kong, that specifically offers human-post-edited ChatGPT MT output as a service.

The author intends to carry out the new translation experiments with ChatGPT Plus as an MT engine and as an automatic post-editor during the next academic year. The plan is to continue to evaluate the results manually and not automatically.

References

1. Farrell, M.: Current evidence of post-editing: differences between post-edited neural machine translation output and human translation revealed through human evaluation. Proposed for: International Conference HiT-IT 2023 - Human-informed Translation and Interpreting Technology. Publication pending (2023).
2. Farrell, M.: Machine Translation Markers in Post-Edited Machine Translation Output. In: Proceedings of the 40th Conference Translating and the Computer, pp. 50–59. AsLing: The International Association for Advancement in Language Technology (2018).
3. Farrell, M.: Raw Output Evaluator, a Freeware Tool for Manually Assessing Raw Outputs from Different Machine Translation Engines. In: Proceedings of the 40th Conference Translating and the Computer, pp. 38–49. AsLing: The International Association for Advancement in Language Technology (2018).
4. Farrell, M.: Do translators use machine translation and if so, how? Results of a survey held among professional translators. Presented at the 44th Conference Translating and the Computer. Preprint pending publication, DOI:10.13140/RG.2.2.33996.69768, (2022).
5. Işım, C., Balcıoğlu, Y. S.: ChatGPT: performance of translate. In Proceedings of 3rd International ACHARAKA Congress on Humanities and Social Sciences (2023).
6. Jiao, W., Wang, W., Huang, J., Wang, X., Tu, Z.: Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. Preprint (2023).

7. Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., Tao, D.: Towards Making the Most of ChatGPT for Machine Translation. Preprint (2023).
8. Toral, A.: Post-editeese: an Exacerbated Translationese. In: Proceedings of Machine Translation Summit XVII: Research Track, pp. 273–281. European Association for Machine Translation, Dublin, Ireland (2019).

Leveraging ChatGPT and Multilingual Knowledge Graph for Automatic Post-Editing

Min Zhang^[0000–0002–9624–6851], Xiaofeng Zhao, Yanqing Zhao, Hao Yang, Xiaosong Qiao, Junhao Zhu, Wenbing Ma, Chang Su, Yilun Liu, Yinglu Li, Minghan Wang, Song Peng, Shimin Tao, and Yanfei Jiang

Huawei Translation Services Center, Beijing, China
{zhangmin186,zhaoxiaofeng14,zhaoyanqing,yanghao30,qiaoxiaosong,zhujunhao,mawenbing,suchang8,liuyilun3,liyinglu,wangminghan,pengsong2,taoshimin,jiangyanfei}@huawei.com

Abstract. Recently, ChatGPT has shown promising results for Machine Translation (MT). However, how to apply ChatGPT for Automatic Post-Editing (APE) remains as an open question. In this paper, we propose a novel zero-shot APE method by leveraging ChatGPT and Multilingual Knowledge Graph (MKG). In this method, we use MKG to find incorrectly translated entities, and then generate APE prompts for ChatGPT with these entities and their correct translations provided in MKG, aiming to have ChatGPT automatically correct the mistranslations. To evaluate our method, we construct two test datasets from WMT19 English-Chinese (En-Zh) and English-German (En-De) news translation shared task. Preliminary experimental results demonstrate that our APE method improves the translation accuracy of entities significantly (+29.1% and +7.3% absolute points for En-Zh and En-De respectively) and achieves a 4.2 BLEU improvement on the En-Zh dataset, showing that our method is effective. However, there is a 7.3 BLEU drop on the En-De dataset, for which we will conduct further research.

Keywords: Multilingual Knowledge Graph, ChatGPT, Automatic Post-Editing, Machine Translation.

1 Introduction

Machine Translation (MT) is widely employed in industrial translation workflows, which is an intermediate step, i.e., generates a raw translation of a given source sentence or document. It is generally followed by a Post-Editing (PE) step to ensure that the quality of the final translation meets required quality standards. Automatic Post-Editing (APE) [20, 2] is an area of research that aims at exploring methods that apply automatic editing operations on an MT output to produce a better translation, thereby reducing human efforts in the translation workflow. A wide range of methods have been proposed for APE, from rule-based approaches [8, 2] to deep learning techniques [7, 25, 1].

Recently, the emergence of ChatGPT¹ has brought remarkable influence on Natural Language Processing (NLP) tasks. ChatGPT is an intelligent chatting machine developed by OpenAI based on InstructGPT [14]. ChatGPT is built upon GPT-3.5 and GPT-4 families of Large Language Models (LLMs) and has been fine-tuned using both supervised and reinforcement learning techniques. ChatGPT possesses diverse abilities of NLP, such as question answering, dialogue generation, code debugging, generation evaluation, and MT [18, 29, 23, 9, 10, 6, 5, 16]. To the best of our knowledge, there is little research on how to apply ChatGPT for APE, although ChatGPT has shown competitive results with commercial translation products (e.g., Google Translator and Microsoft Translator) [6, 5, 16].

During translation, entities in a sentence play an important role, and their correct translation can heavily affect the whole translation quality of this sentence. Due to the importance of entities, various methods are proposed to improve their translation or perform Quality Estimation (QE) with Knowledge Graph (KG) [19, 11, 13, 27, 28, 26, 4, 24]. However, to the best of our knowledge, none of these works are for APE.

In this paper, we propose a novel zero-shot APE method by leveraging ChatGPT and Multilingual Knowledge Graph (MKG). First, we find the incorrectly translated entities in source sentences by MKG. Second, we generate APE prompts for ChatGPT with these entities and their correct translations provided in MKG. Finally, we request ChatGPT to output APE results based on our prompts. Since no APE training data is used, our APE method is *zero-shot*, which is different from the above mentioned methods in the APE field. To evaluate this method, we conduct preliminary experiments on two language pair (En-Zh and En-De) datasets from WMT19 news translation shared task [12]. Experimental results show that APE method significantly improves the translation accuracy of entities on the two datasets (+29.1% and +7.3% absolute points for En-Zh and En-De respectively). Meanwhile, our APE method achieves a 4.2 BLEU improvement on the En-Zh dataset but a 7.2 BLEU drop on the En-De dataset. This suggests further research is needed.

The main contributions of this paper are as follows:

- To the best of our knowledge, we are the first to propose a novel zero-shot APE method by leveraging ChatGPT and MKG.
- Preliminary experimental results show that our APE method is very effective to improve the translation accuracy of entities.

2 Related Work

2.1 APE

APE can be seen as a *monolingual* translation task [21], and the same MT technology can be used to develop APE systems. Unlike MT, where the system

¹ <https://chat.openai.com>

learns bilingual translations from source and target pairs (src, tgt), APE learns to correct errors from MT text and human PEs (mt, pe) or from triplets (src, mt, pe) to leverage the source context too. Most of the current state-of-the-art (SOTA) APE systems are built using triplets [20, 2].

The development of APE has gone through rule-based paradigm, phrase-based paradigm to neural paradigm [3]. Rule-based systems generally have precise PE rules as they are manually written. However, these hand-crafted rules are insufficient to capture all possible scenarios. Phrase-based technology has been used in most work since the beginning of APE. It provides faster training and decoding capabilities and can learn efficiently from small datasets. Neural technology has emerged as a stronger alternative to phrase-based methods, achieving SOTA performance in APE. However, since current APE systems use the same network architecture as MT Transformer [22], much less training data is used, so the improvement of the current SOTA APE systems on the basis of MT is not significant enough [20, 2].

Different from the current APE systems, ChatGPT adopts a different network architecture from MT and uses large-scale massive training data, so it is likely to have a better effect for APE.

2.2 Knowledge Graph for MT

With the help of KG, Shi et al. built and formulated a semantic space to connect the source and target languages, applied it to the sequence-to-sequence framework, and proposed a Knowledge-Based Semantic Embedding method [19]. Lu et al. utilized the entity relations in KG as constraints to enhance the connections between the source words and their translations [11]. Under the hypothesis that KG could enhance the semantic feature extraction of neural models, Moussallem et al. proposed two strategies for incorporating KG into neural models without modifying the neural network architectures [13]. Zhao et al. not only proposed a multi-task learning method on sub-entity granularity for MT task and knowledge reasoning task [27], but also designed a novel KG enhanced Neural Machine Translation (NMT) method (i.e., transforming the source and target KGs into a unified semantic space) [28]. To apply the entity pairs of MKG, Zhang et al. proposed a data augmentation strategy for NMT [26].

In a word, all the above approaches are about how to introduce KG into neural networks for MT, not for APE. In this paper, we utilize MKG to generate APE prompts for ChatGPT.

2.3 ChatGPT for MT

With ChatGPT showing remarkable capabilities in various NLP tasks, research on ChatGPT for MT has sprung up [6, 5, 16].

Jiao et al. provided a preliminary evaluation of ChatGPT for MT, including translation prompt, multilingual translation, and translation robustness [6].

Hendy et al. presented a comprehensive evaluation of ChatGPT for MT, covering various aspects such as quality of ChatGPT in comparison with state-of-the-art research and commercial systems, effect of prompting strategies, robustness towards domain shifts and document-level translation [5]. These studies show that ChatGPT does not perform as well as commercial translation products on low-resource languages or specific domains. Peng et al. proposed two simple yet effective prompts (task-specific and domain-specific prompts) to mitigate these issues [16].

The above works discuss the translation ability of ChatGPT, but do not involve its APE ability. In this paper, we conduct a preliminary study on the APE capability of ChatGPT.

3 Method

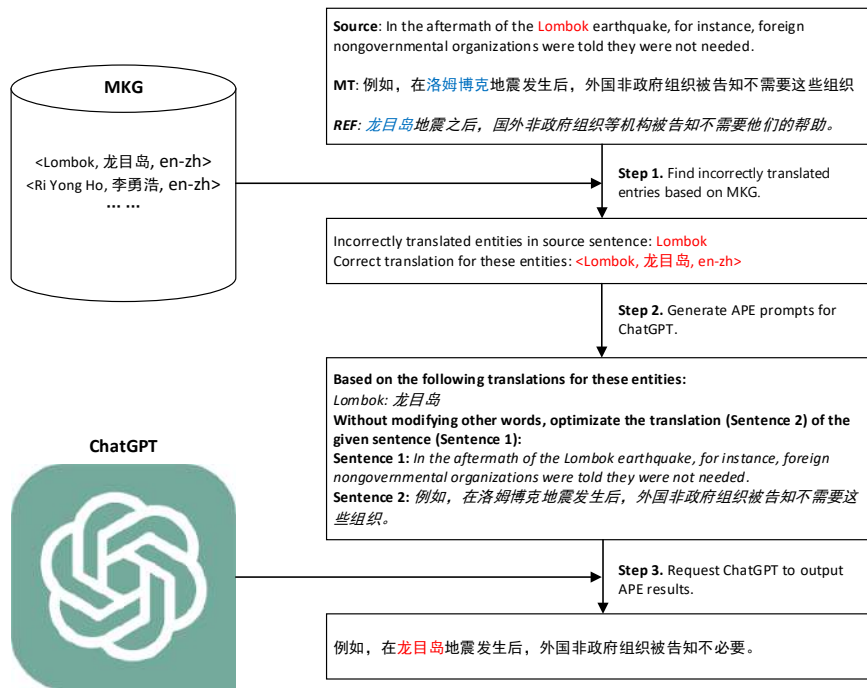


Fig. 1. An En-Zh example for the pipeline of our APE method, where the entity “Lombok” in the source sentence should be translated to “龙目岛” but is wrongly translated to “洛姆博克” in MT.

The pipeline of our APE method by leveraging ChatGPT and MKG is illustrated in Fig. 1, which consists of three steps. In this paper, MKG is composed of triplets as $\langle \textit{source entity}, \textit{target entity}, \textit{language pair} \rangle$. For example, the triplet “ $\langle \textit{Lombok}, \textit{龙目岛}, \textit{en-zh} \rangle$ ” in Fig. 1 indicates the Chinese translation of English entity *Lombok* is Chinese entity 龙目岛. Details of the three steps are described as follow.

In Step 1, given the source sentence and its translation, we first find all the entities in the source sentence, which are contained in the source entities of MKG. In Fig. 1, the entity “*Lombok*” is found. Then we check whether the correct translations of these entities appear in the translation. If not, the corresponding triplets of these entities in the MKG is outputted. In Fig. 1, the correct translation “龙目岛” of entity “*Lombok*” does not appear in the translation, so the triplet “ $\langle \textit{Lombok}, \textit{龙目岛}, \textit{en-zh} \rangle$ ” is outputted for Step 2. It should be noted that the APE process will be terminated if the output of Step 1 is empty (i.e., the entities in the source sentence are all correctly translated in its translation according to MKG).

In Step 2, we generate APE prompts for ChatGPT based on the triplets outputted in Step 1. As shown in Fig. 1, the APE prompts contain the fixed content (in bold), the triplets, and the source sentence and its translation.

In Step 3, we request ChatGPT to output the APE results with the prompts generated in Step 2. In Fig. 1, the APE result from ChatGPT successfully fixes the mistranslation of the entity “*Lombok*”.

4 Experiments

4.1 Experimental Settings

We provide a brief description of the experiment settings, which mainly include the used models, test datasets, MKG and evaluation metrics.

Models The *gpt-3.5-turbo* model which powers the ChatGPT is used in our APE method.

Data To evaluate the APE method, we construct two test datasets from WMT19 news translation shared task [12], where the two best system translations (KSAI and Facebook-FAIR) on the En-Zh and En-De language pairs are selected. The MKG is built on the entity annotation results of source sentences and reference sentences in WMT19 [4], which contains 4,726 and 4,115 triplets for En-Zh and En-De respectively. Since there are no entity translation errors in the translations of sentences, we filter these sentences from the two selected datasets (1,624 out of 1,997 sentences for En-Zh, and 1,704 out of 1,997 sentences for En-De), and finally the sizes of the two constructed test datasets are 375 for En-Zh and 295 for En-De respectively, which indicates less than 20% of sentences having entity translation errors. In this section, we try to correct the entity translation errors in these sentences via our APE method.

Table 1. Results of the APE method on the En-Zh and En-De datasets from WMT19.

System	En⇒Zh		En⇒De	
	BLEU	ACC	BLEU	ACC
WMT19 Best MT	32.78	0.523	33.45	0.634
APE (MKG+ChatGPT)	36.98	0.814	26.25	0.707

Table 2. Results of MT by ChatGPT on the En-Zh and En-De datasets from WMT19.

System	En⇒Zh		En⇒De	
	BLEU	ACC	BLEU	ACC
WMT19 Best MT	32.78	0.523	33.45	0.634
MT (ChatGPT)	33.94	0.657	32.70	0.625

Table 3. Two Cases of the APE method from the En-Zh test dataset.

ID	Src	MT	Entities	APE	Ref
1	Respiratory irritation continues to be reported in Pinellas, Manatee, Sarasota, Lee, and Collier counties	在皮涅拉斯州、马纳提州、萨拉索塔州、李州和科利尔州，仍有呼吸道刺激的报告。	<Manatee,马纳提县, en-zh> <Collier,科利尔县, en-zh>	在皮涅拉斯县、马纳提县、萨拉索塔县、李县和科利尔县，仍有呼吸道刺激的报告。	皮内拉斯县、马纳提县、萨拉索塔县、李伊县和科利尔县的报告显示仍存在呼吸道刺激。
2	A man has been shot multiple times with an air rifle as he walked home from a night out.	一名男子在一个晚上外出回家的路上被一支气枪击中多次。	<rifle,步枪, en-zh>	一名男子在一个晚上外出回家的路上被一支步枪击中多次。	一名男子于晚间外出回家途中，遭气步枪多次射击。

Metrics We adopt the commonly used BLEU score [15] as our primary metric, which is calculated by the toolkit SacreBLEU [17]. Additionally, we report the translation accuracy (ACC) of entities in the APE results.

4.2 Experimental Results

Table 1 shows the metric results of our APE method on the two constructed test datasets, where the metric values of the best system translations selected from WMT19 are provided for comparison.

From Table 1, it can be seen that our APE method significantly improves the translation accuracy of entities on the two datasets (+29.1% and +7.3% absolute points for En-Zh and En-De respectively) and the BLEU score of the En-Zh translations from WMT19 by 4.2, which indicates that the APE method is effective. However, the APE method has a drop by 7.2 in BLEU on the En-De

translations from WMT19. This means that our APE method is not robust for En-De translations and further research is needed.

In addition, we report the metric values of translation results by ChatGPT in Table 2, where two translation prompts “*Please translate the following sentence into Chinese: <sentence>*” and “*Please translate the following sentence into German: <sentence>*” are used to get the translation results of ChatGPT for En-Zh and En-De respectively. It can be seen that the translations by ChatGPT are competitive with the best MT results from WMT19 (1.16 BLEU higher on En-Zh, and 0.75 BLEU lower on En-De). And the translation accuracy of entities on En-Zh is significantly higher, possibly due to the massive training corpus for ChatGPT.

4.3 Case Studies

In this section, we provide two cases of the APE method on the En-Zh test dataset as illustrated in Table 3, where the “*Entities*” column denotes the incorrectly translated entities in *Src* and their correct translations in MKG. These entities and their translations are highlighted in different colors in *Src*, *MT*, *APE* and *Ref*.

From Case 1, it can be seen that the two mistranslated entities “*Manatee*” and “*Collier*” are correctly translated by our APE method. However, the mistranslated entities “*Pinellas*” and “*Lee*”, which are not covered in MKG, are not fixed. This suggests that the coverage of MKG is extremely important for our APE method.

From Case 2, although the APE method correctly fixes the mistranslated entity “*rifle*”, it causes a missing translation problem, i.e., the translation of word “*air*” is missed in APE. This suggests that the APE method is not robust enough and further research is needed.

5 Conclusion

In this paper, a novel zero-shot APE method is proposed by leveraging ChatGPT and MKG. With incorrectly translated entities extracted based on MKG, APE prompts are generated, and ChatGPT is requested to output the APE results by following the prompts. Two datasets (En-Zh and En-De) are constructed from WMT19 for experiments. Experimental results demonstrate that the APE method gets a 4.2 BLEU improvement on the En-Zh dataset but a 7.2 BLEU drop on the En-De dataset, although the translation accuracy of entities on both datasets is significantly improved (+29.1% and +7.3% absolute points of ACC for En-Zh and En-De respectively). This suggests that the APE method is not robust enough and further research is needed.

References

1. Bhattacharyya, P., Chatterjee, R., Freitag, M., Kanojia, D., Negri, M., Turchi, M.: Findings of the WMT 2022 shared task on automatic post-editing. In: Proceedings

- of the Seventh Conference on Machine Translation (WMT). pp. 109–117. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022), <https://aclanthology.org/2022.wmt-1.5>
2. do Carmo, F., Shterionov, D., Moorkens, J., Wagner, J., Hossari, M., Paquin, E., Schmidtke, D., Groves, D., Way, A.: A review of the state-of-the-art in automatic post-editing. *Machine Translation* **35**(2), 101–143 (jun 2021), <https://doi.org/10.1007/s10590-020-09252-y>
 3. Chatterjee, R.: Automatic post-editing for machine translation. *ArXiv abs/1910.08592* (2019)
 4. Gekhman, Z., Aharoni, R., Beryozkin, G., Freitag, M., Macherey, W.: KoBE: Knowledge-based machine translation evaluation. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 3200–3207. Association for Computational Linguistics, Online (Nov 2020), <https://aclanthology.org/2020.findings-emnlp.287>
 5. Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y.J., Afify, M., Awadalla, H.H.: How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210* (2023)
 6. Jiao, W., Wang, W., tse Huang, J., Wang, X., Tu, Z.: Is chatgpt a good translator? a preliminary study. In: *ArXiv* (2023)
 7. Junczys-Dowmunt, M., Grundkiewicz, R.: An exploration of neural sequence-to-sequence architectures for automatic post-editing. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 120–129. Asian Federation of Natural Language Processing, Taipei, Taiwan (Nov 2017), <https://www.aclweb.org/anthology/I17-1013>
 8. Knight, K., Chander, I.: Automated postediting of documents. In: Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence. p. 779–784. AAAI’94, AAAI Press (1994)
 9. Kocmi, T., Federmann, C.: Large language models are state-of-the-art evaluators of translation quality. *ArXiv abs/2302.14520* (2023)
 10. Lu, Q., Qiu, B., Ding, L., Xie, L., Tao, D.: Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint* (2023), <https://arxiv.org/pdf/2303.13809.pdf>
 11. Lu, Y., Zhang, J., Zong, C.: Exploiting knowledge graph in neural machine translation. In: Chen, J., Zhang, J. (eds.) *Machine Translation*. pp. 27–38. Springer Singapore, Singapore (2019)
 12. Ma, Q., Wei, J., Bojar, O., Graham, Y.: Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). pp. 62–90. Association for Computational Linguistics, Florence, Italy (Aug 2019), <https://aclanthology.org/W19-5302>
 13. Moussallem, D., Ngonga Ngomo, A.C., Buitelaar, P., Arcan, M.: Utilizing knowledge graphs for neural machine translation augmentation. In: Proceedings of the 10th International Conference on Knowledge Capture. p. 139–146. K-CAP ’19, Association for Computing Machinery, New York, NY, USA (2019), <https://doi.org/10.1145/3360901.3364423>
 14. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) *Advances in Neural*

- Information Processing Systems. vol. 35, pp. 27730–27744. Curran Associates, Inc. (2022)
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002), <https://aclanthology.org/P02-1040>
 16. Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., Tao, D.: Towards making the most of chatgpt for machine translation. arxiv preprint (2023), <https://arxiv.org/abs/2303.13780>
 17. Post, M.: A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers. pp. 186–191. Association for Computational Linguistics, Belgium, Brussels (Oct 2018), <https://www.aclweb.org/anthology/W18-6319>
 18. Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., Yang, D.: Is chatgpt a general-purpose natural language processing task solver? ArXiv **abs/2302.06476** (2023)
 19. Shi, C., Liu, S., Ren, S., Feng, S., Li, M., Zhou, M., Sun, X., Wang, H.: Knowledge-based semantic embedding for machine translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2245–2254. Association for Computational Linguistics, Berlin, Germany (Aug 2016), <https://aclanthology.org/P16-1212>
 20. Shterionov, D., do Carmo, F., Moorkens, J., Hossari, M., Wagner, J., Paquin, E., Schmidtko, D., Groves, D., Way, A.: A roadmap to neural automatic post-editing: an empirical approach. *Machine Translation* **34**, 67 – 96 (2020)
 21. Simard, M., Goutte, C., Isabelle, P.: Statistical phrase-based post-editing. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference. pp. 508–515. Association for Computational Linguistics, Rochester, New York (Apr 2007), <https://aclanthology.org/N07-1064>
 22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6000–6010. NIPS’17, Curran Associates Inc., Red Hook, NY, USA (2017)
 23. Wang, J., Liang, Y., Meng, F., Li, Z., Qu, J., Zhou, J.: Cross-Lingual Summarization via ChatGPT. arXiv.org (Feb 2023)
 24. Wu, Z., Zhang, M., Zhu, M., Li, Y., Zhu, T., Yang, H., Peng, S., Qin, Y.: Kgbertscore: Incorporating knowledge graph into bertscore for reference-free machine translation evaluation. In: Proceedings of the 11th International Joint Conference on Knowledge Graphs. p. 121–125. IJCKG ’22, Association for Computing Machinery, New York, NY, USA (2023), <https://doi.org/10.1145/3579051.3579065>
 25. Yang, H., Wang, M., Wei, D., Shang, H., Guo, J., Li, Z., Lei, L., Qin, Y., Tao, S., Sun, S., Chen, Y.: HW-TSC’s participation at WMT 2020 automatic post editing shared task. In: Proceedings of the Fifth Conference on Machine Translation. pp. 797–802. Association for Computational Linguistics, Online (Nov 2020), <https://aclanthology.org/2020.wmt-1.85>
 26. Zhang, M., Peng, S., Yang, H., Zhao, Y., Qiao, X., Zhu, J., Tao, S., Qin, Y., Jiang, Y.: Entityrank: Unsupervised mining of bilingual named entity pairs from parallel corpora for neural machine translation. In: 2022 IEEE International Conference on Big Data (Big Data). pp. 3708–3713 (2022). <https://doi.org/10.1109/BigData55660.2022.10021032>

27. Zhao, Y., Xiang, L., Zhu, J., Zhang, J., Zhou, Y., Zong, C.: Knowledge graph enhanced neural machine translation via multi-task learning on sub-entity granularity. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 4495–4505. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020), <https://www.aclweb.org/anthology/2020.coling-main.397>
28. Zhao, Y., Zhang, J., Zhou, Y., Zong, C.: Knowledge graphs enhanced neural machine translation. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. pp. 4039–4045. International Joint Conferences on Artificial Intelligence Organization (7 2020), <https://doi.org/10.24963/ijcai.2020/559>, main track
29. Zhong, Q., Ding, L., Liu, J., Du, B., Tao, D.: Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. arXiv preprint (2023). <https://doi.org/10.48550/arXiv.2302.10198>

The Proper Place of Men and Machines - Updated

Elliott Macklovitch

Université de Montréal, Montréal, Canada
firstname.familyname@umontreal.ca

Abstract. The central claim of Martin Kay’s famous article “The Proper Place of Men and Machines in Language Translation” is that one cannot automate what one does not fully understand – in this case, translation. In light of the impressive quality of the translations produced by current neural machine translation systems (henceforth, NMT), we return to that claim and examine whether it still obtains for NMT, which implements a variety of distributional semantics. Acknowledging the obvious, i.e. that machines clearly do not understand in the same way that humans do, we contend that NMT translations are indeed meaning-based. In support of that contention, we point to the success of certain NMT systems in achieving so-called zero-shot translation between languages where no explicit training data are available to them. We conclude by suggesting that these systems now require us to invert the proper roles of men and machines in language translation.

Keywords: neural machine translation, Martin Kay, language understanding.

1 Introduction

In 1980, Martin Kay, the late, great computational linguist, published his celebrated paper ‘The Proper Place of Men and Machines in Language Translation’.¹ Although the paper is famous today, it should be noted that it was not particularly well received at the time. In particular, it did little to slow or hamper the development of a host of machine translation (henceforth, MT) projects, all of which had as their aim fully automatic, high-quality MT. These included many commercial systems, such as Logos, Sys-tran, METAL and a host of Japanese projects, as well many university or otherwise publicly funded projects like EUROTRA. And while some may have claimed to serve merely as machine aids to human translators, in assuming the full responsibility for the translation process and relying on the human solely to clean up the machine output, they all fell within the purview of Kay’s critique.

‘The Proper Place’ is a remarkable paper for many reasons, not the least of which is the verve and sting of Kay’s prose. To cite just one example:

“There was a long period – for all I know, it is not yet over – in which the following comedy was acted out nightly in the bowels of an American government office with the aim of rendering foreign texts into English. Passages of innocent prose on which it was desired to effect this

¹ The paper first appeared as a Xerox research report in 1980 and was later republished in the *Machine Translation* journal vol. 12 (1997), as well as in the collection *Readings in Machine Translation*, MIT Press (2003). The page numbers of citations included here are those of the *Machine Translation* version.

delicate and complex operation were subjected to a process of vivisection at the hands of an uncomprehending electronic monster that transformed them into stammering streams of verbal wreckage. These were then placed into only slightly more gentle hands for repair.” (p. 5)

2 Kay’s Credo

In ‘The Proper Place’, Kay approvingly cites the previous work of Yehoshua Bar-Hillel, who, as far back as the early 1950’s, had convincingly argued for the non-feasibility of fully automatic, high-quality machine translation (coining in the process the well-known acronym FAHQT); see Bar-Hillel [1]. Bar-Hillel’s well-known thought experiment is based on the need to access and reason over unpredictable and potentially infinite amounts of real-world knowledge, even to translate a sentence as innocuous as “the box is in the pen”, a requirement he felt was and would remain absolutely unattainable for a translating machine. However, this did not prevent him from supporting machine translation as an excellent vehicle for fundamental linguistic research. What he deplored was the misguided promise that MT could replace human translators, or even serve to enhance their productivity in the short or medium term.

Kay fully agrees with this position, but the principal argument he advances against FAHQT in ‘The Proper Place’ comes from a different angle – that of a computer scientist.² Kay’s explanation for “the stammering streams of verbal wreckage” produced by the MT systems of his day is quite simple: “it happens when the attempt is made to mechanize the non-mechanical or something whose mechanistic substructure science has not yet revealed. In other words, it happens when we attempt to use computers to do something we do not fully understand. History provides no better example of the improper use of computers than machine translation.” (p.4) And later in the paper: “There is a great deal that computer scientists and linguists could contribute to the practical problem of producing translations, but, in their own interests, as well as those of their customers, they should never be asked to provide an engineering solution to a problem they only dimly understand.” (p.5)

What Kay is clearly implying in these statements is that our current understanding of translation, and perhaps more generally, our understanding of how human language works, is very partial and definitely insufficient to allow us to provide coded instructions to a machine on how to translate texts between two natural languages. For many years, in my own classes on translation technology, I used to refer to Kay’s argument in an effort to be explain to my students the poor quality of most machine translation output. Simplifying somewhat, what I would tell them was this: you can’t automate what you don’t understand. Or put another way: if you do attempt to automate what you don’t understand – in this case, translation – then this is the kind of output you obtain.

² In one of his many postings, Kay served as Chair of the Department of Computer Science at the University of California at Irvine.

3 That was then and this is...

2014 is the year that is generally cited as marking the advent of neural machine translation (NMT), the seminal articles often mentioned being Bahdenau et al. [2], Cho et al. [3], and Sutskever et al. [4]. So NMT has been with us for nearly ten years now, and it definitely has continued to improve over that period. Indeed, the best of today's NMT systems have reached a level that is so impressive that I have no qualms in asserting – and not just for the sake of the argument that follows – that FAHQT has finally been achieved. At first blush, this may sound like an overblown claim, and so we need to carefully consider what it does *not* entail. It does not mean that today's NMT systems always generate perfect translations; on the contrary, it is well known that they do occasionally produce translations that are incorrect in one way another, and sometimes bafflingly so. (But then again, what human translator can honestly claim to never making mistakes?) Nor is it to assert that the error-free translations they most often provide cannot be improved by a human revisor. (And again, the same is true for human translators.) Granting these provisos, it must be admitted that the output of today's best NMT systems³ bears absolutely no resemblance to the output provided by previous generations of machine translation, even the relatively recent statistical MT systems. NMT translations are almost always grammatically correct and idiomatic; and in most cases, they do convey the essential meaning of the source text being translated. That is precisely the goal that Bar-Hillel established for machine translation when he coined the term FAHQT.

4 Translation and Understanding

Here is another of Kay's bold quotations from 'The Proper Place':

"To translate is to re-express in a second language what has been understood by reading a text. Any purported solution to the problem that does not involve understanding in sense is, at best, ad hoc and therefore subject to the linguistic objections already alluded to." (p.7)

As argued in the previous section, I will assume that current NMT systems have indeed achieved the long-elusive goal of FAHQT. It then seems to me that one of two conclusions must necessarily follow: either such systems do indeed incorporate the kind of "understanding in sense" that Kay is alluding to in the above paragraph; or if they do not, then Kay's basic tenet is incorrect, since these systems continue to demonstrate that they can adequately translate what they do not understand.

Let me begin by stating the obvious: understanding is not a simple, monolithic notion that would allow one to unequivocally assert that, yes, NMT systems do understand the texts that they process; or no, they do not. Rather, understanding is a murky, loosely

³ In speaking of the *best* NMT systems, I am referring to those that have been trained on very large quantities of high-quality data. This is the case for English and French, the language pair that I work with. NMT output on language pairs for which the training data is insufficient will necessarily be less good.

defined concept that surely comprises different levels and admits of many different definitions. Take, for example, my own understanding of the artificial neural networks that underpin all NMT systems. For a translator and an old-school linguist⁴ like myself, these systems seem prodigiously complex and are exceedingly difficult to fully grasp. That said, it appears that the vectorized embeddings that play such a central role in NMT can be traced back to the linguistic theory known as distributional semantics, whose central postulate was famously encapsulated by JR Firth as “you shall know a word by the company it keeps”; see Firth [5]. These word embeddings certainly manage to encode a great deal of information, semantic and otherwise, about each lexical unit in the vocabulary, including many (or all?) of the words it tends to cooccur with. From that point in the neural architecture, however, things become rather mysterious for me, as these lexical embeddings are then merged in hidden layers into sentence embeddings and ultimately converted into numerical representations that are projected into an abstract multi-dimensional space.

The overall encoder-decoder architecture of NMT systems is often said to be much simpler than that of the previous generation of statistical MT systems; nevertheless, the inner workings of NMT remain opaque to most users.⁵ What are we to make of these vectorized sentence embeddings produced by the encoder of a neural MT sentence? One possible approach that I personally find helpful refers back to the famous Vauquois triangle, which the French MT pioneer first proposed in 1968; see Vauquois [7].

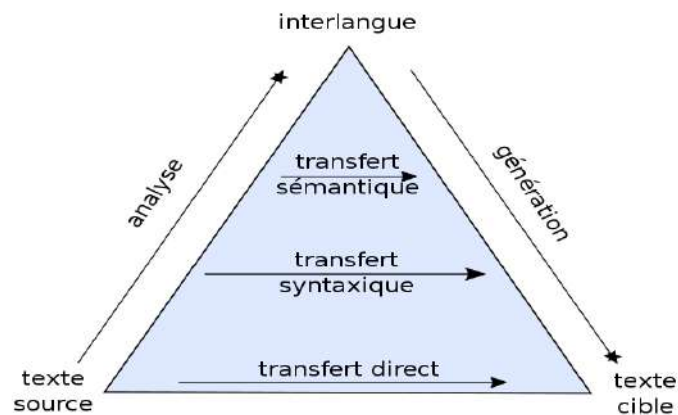


Fig. 1. The Vauquois triangle

⁵ An old-school *computational* linguist, moreover, who spent the greater of part of his professional career working on machine translation R&D projects.

⁵ Including some of the AI specialists who develop the systems! See Lee et al. [6] for their comments on the ‘hallucinatory’ output of their own NMT system.

In suggesting this schema, Vauquois’ central point was this: the deeper the analysis that an MT system carries out on the source text, the less work its transfer component will have to do. As we progress up the triangle from first-generation direct MT systems, which conduct little or no analysis, to second-generation systems, which perform a syntactico-semantic analysis of the input, the number of necessary transfer operations is found to decrease.⁶ At the very tip of the triangle sits something conceived as a universal interlingua. Here, transfer disappears entirely; no transformations whatsoever are required to pass from one language to another. The classical examples often cited are Arabic numerals and chemical formulas: from NaCl, one can directly generate the linguistic expression ‘sodium chloride’ in English, ‘chlorure de sodium’ in French, and so on in other languages, without any further operations required.

Now one might be tempted to consider the vectorized embeddings produced by NMT systems as more elaborate instantiation of this interlingua. And indeed, there are certain AI researchers who make this claim quite explicitly; see, for example Lu et al. [8] and Escolano et al. [9]. The latter propose a multilingual MT systems that uses “multiple encoders and decoders for each language, sharing a *common intermediate representation*.” The former describe “an attentional neural interlingua that receives language-specific encoder embeddings which are *agnostic to the source and target language*.” (my emphasis in both cases) On the other hand, in Google’s large-scale work on massively multilingual neural MT (see Arivazhagen et al. [10]), no mention is made of an interlingua. And when Angela Fan, the team leader of Meta’s No Language Left Behind project, is directly asked the question about the status of their system’s intermediate representations,⁷ she demurs, recognizing that the promise of multilingual MT has always been “some kind of multilingual space”, but refuses to commit, saying instead that this is still an active area of investigation. (See her Youtube interview [11]; also [15] for a complete description of the NLLB project).

Notice, however, that the Vauquois triangle also includes a level just below the interlingua pinnacle, which is referred to as semantic transfer. Vauquois did not have much to say about this level of representation at the time, but in subsequent decades numerous formalisms have been advanced as candidates for semantic representations, although, as far as I know, few have been proposed within the framework of machine translation as providing the basis for deep linguistic transfer between languages.⁸ Here, I would like to (timidly) suggest that the vectorized numerical representations generated by neural machine translation systems are better viewed as constituting the kind of semantic representations that Vauquois had in mind, rather than as instantiations of a universal interlingua. That these vectorized embeddings are (largely) semantic in nature is only to be expected from an approach inspired by distributional semantics. Moreover, they have allowed for some relative success on zero-shot translation, i.e. the ability to

⁶ It is no accident, in other words, that Vauquois’ schema is shaped in the form of a triangle and not as a rectangle, for example.

⁷ The question posed is incorrectly formulated in terms of Chomsky’s universal grammar, but Angela Fan correctly interprets it to refer to a universal interlingua.

⁸ One exception that comes to mind is Lexical Functional Grammar, which has been used on a few machine translation projects. With how much success, I cannot say.

produce translations between language pairs for which no explicit training data is available. (Indeed, it is hard to imagine what else except for a semantic representation could possibly allow for translation between languages for which no explicit training data has been employed.) However, unlike the simple interlingual examples cited above, these zero-shot translations are often imperfect, meaning that further transfer operations (of some sort) would be required to transform them into target output that is fully meaning preserving. And that alone, in my view, is sufficient to disqualify them as being interlingual.⁹

Let us now return to Kay’s basic credo and once again ask the question: Does an NMT system understand the texts it is attempting to translate? Obviously, it does not achieve the same kind of understanding that human translators have when they translate a sentence. The system has no notion of the objects, processes and events that the words and sentences refer to in the real world, outside the texts.¹⁰ On the other hand, it does not seem to me unreasonable to claim that it does understand something very fundamental about translation. In proposing target sentence y as a translation of source sentence x , the system is implicitly making the claim that it understands the two sentences to mean the same thing. To take one simple example: an MT system needn’t comprehend what a ‘free-falling body’ refers to in the real world in order to know that the term is translated as ‘un corps en chute libre’ in French. Of course, it could be said that all machine translation systems have always been making this same implicit claim. True enough; but only NMT systems have managed to achieve a level of translation success that impels us to take this claim seriously.

We have been arguing that NMT systems do have a certain understanding of the texts they process and that the translations they produce are indeed meaning-based. That this understanding is not the same as that of a human translator is obvious enough,¹¹ but why should this matter? For years, we attempted to program the machine to emulate what we thought was the manner in which human translators operated, with very limited success. It was only when the rule-based, expert system approach was abandoned in favour of applying machine learning techniques to very large corpora of translated text that MT systems slowly began to improve.¹² And it was only when artificial neural networks were applied to that same task that machine translation output began to improve dramatically.

⁹ It is not sufficient, in other words, that these representations be “agnostic” between the source and target language; they also have to be adequate to directly generate a fully correct translation. For another take on the question, see do Carmo [12], his lecture on certain MT myths. While accepting the interlingual thesis, he is more skeptical of the claims made for zero-shot translation.

¹⁰ Piantadosi & Hill [13] convincingly argue that this does not prevent the representations learned by large language models from encoding important semantic information.

¹¹ Alan Melby is another who has argued that MT systems have no understanding of language, but merely “manipulate words mechanically”. See Melby and Kurz [14].

¹² The analogy that immediately comes to mind is human flight: it was only when people stopped flapping their arms like birds that human flight finally got off the ground.

That is the upside of the argument; but there is also a downside to the fact that NMT systems do not operate in any way similar to human translators. As mentioned above, the internal operations of these systems are excessively complex and difficult for those who use them to comprehend. For most working translators,¹³ an NMT system is very much a black box, one that we are still learning to work with. In previous generations of MT, translators enjoyed a modicum of control which allowed them, for example, to correct a dictionary entry and thereby alter the system's output. It is not yet clear how or to what extent users can modify the behavior of today's NMT systems, short of wholesale retraining. And retraining these data-hungry behemoths is no simple matter.

5 The place of the machine, reassigned

The second half of Kay's 'Proper Place' article is devoted to a concrete proposal for what he considers to be a more reasonable way of using computers to help working translators cope with the ever-increasing demand for their services. He calls his proposal a translator's amanuensis; on Pierre Isabelle's team at the CITI, we called a very similar project a translator's workstation.¹⁴ In both cases, what was being proposed was basically a multilingual word processor supplemented with a number of independent programs designed to assist the translator with various ancillary tasks, e.g., file format conversion, a personal glossary, spell checking, etc. Of course, all this sounds elementary today, but it has to be recalled that at the time of Kay's paper, the first popular personal computer (the IBM PC) had not yet been launched and very few translators had any experience working on a computer at all.

A key feature of Kay's proposal was its incremental nature:

"I want to advocate a view of the problem in which machines are gradually, almost imperceptibly, allowed to take over certain functions in the overall translation process. First they will take over functions not essentially related to translation. Then, little by little, they will approach translation itself. The keynote will be modesty. At each stage, we will do only what we know we can do reliably. Little steps for little feet!" (p.13)

Given Kay's negative assessment of machine translation, it is somewhat surprising to find that MT was not entirely banished from his amanuensis. But actually, the real target of Kay's attack is less machine translation itself than the manner in which these systems were employed at the time and the subordinate role that was left to the human translator. The standard *modus operandi* corresponded to what Kay colorfully described in the quotation given on the first page above: texts were first processed by the MT system (which invariably ran on a mainframe computer) and then passed on to a translator for correction. In Kay's amanuensis, on the other hand, it is the translator who firmly sits in the driver's seat; they are in complete control of the translation process

¹³ As opposed to the AI specialists who develop these systems; and even they appear to struggle with the systems' opacity. See Bau et al. [16] for the description of a study that aims to control the artificial neurons that determine a particular NMT output.

¹⁴ On the CITI's workstation project, see Macklovitch [17].

and may, if so desired, request a machine translation of a portion of text, which they can then accept, post-edit or decide to ignore.

As we mentioned above, MT system output did not substantially improve for many years (if not decades) following the publication of Kay's article. Hence, it was only natural that this secondary, optional recourse to machine translation which Kay ascribed to MT remain in effect, at least on those projects where translators had a say. Even today, within many translation environment tools that incorporate both translation memory (TM) and machine translation, priority is routinely given to the former over the latter. Concretely, this usually means that if a match for a given segment is found within the TM database, it is the one that is inserted by default into the system's editor, on the grounds that the translation memory contains human translations, which are assumed to be necessarily superior to those generated by an MT system.

Given the dramatic improvement in the quality of the translations generated by current NMT systems, I am not convinced that this division of labour between MT and TM remains valid today. At the very least, I believe that translators should always have access to the NMT output, alongside the TM output. Nor would I be surprised to learn that in a significant number of cases, translators choose to adopt or post-edit the MT output in preference to that retrieved from the memory.¹⁵

Yet, as previously mentioned, even the best NMT systems still occasionally produce erroneous translations, sometimes in the form of omitted content, less often in the form of wildly egregious (but grammatical) output. Because these errors remain by-and-large unpredictable, a qualified human translator will necessarily be required to revise all NMT translations that are either destined for wide dissemination or include content that could compromise security or potentially pose a danger.¹⁶ Why not just a proof-reader who is a native speaker of the target language? Because, paradoxically, it is much more difficult to detect the occasional semantic slip-up in the perfectly fluid output of NMT systems than it was to spot the often ungrammatical output that leapt off the page and demanded correction in the output of previous generations of MT. As the adoption of neural MT continues to grow, more and more translators will find themselves recruited to perform this kind of MT revision or, in the case of texts intended for publication, fine-grained MT post-editing. We may or may not like it, but for many of us, I am convinced that this is destined to become our proper place in the translation process in the coming years.

¹⁵ Particularly since the discrete segments stored in TM do not take the larger extra-sentential context into account, something that NMT systems are beginning to do. See for example Bawden et al. [18].

¹⁶ This point too was made by Kay in 'The Proper Place', where he argues against the often evoked statistical defense of MT, stating: "An algorithm that works most of the time is, in fact, of very little use unless there is an automatic way of deciding when it is and when it is not working." (p.10)

References

1. Bar-Hillel, Y.: The Present Status of Automatic Translation of Languages. *Advances in Computers* (1), 91–163 (1960).
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. <https://doi.org/10.48550/arXiv.1409.0473> (2014).
3. Cho, Kyunghyun, et al. "On the properties of neural machine translation: Encoder-decoder approaches." arXiv preprint arXiv:1409.1259 (2014)
4. Sutskever, I., Vinyals, O., and Le, Q. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS 2014)*.
5. Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pp. 1-32. Oxford: Philological Society. Reprinted in F.R. Palmer (ed.), *Selected Papers of J.R. Firth 1952-1959*, London: Longman (1968).
6. Lee, K., Firat, O., Agarwal, A., Fannjiang, C., Sussillo, D. Hallucinations in Neural Machine Translation. In: *Conference on Neural Information Processing Systems, Montréal, Canada. (2018)*
7. Vauquois, Bernard. A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In : *Ifip congress (2)*. 1968. p. 1114-1122.
8. Lu, Y., Keung, P., Ladhak, F., Bhardwaj, V., Zhang, S., Sun, J. A neural interlingua for multilingual machine translation. arXiv: 1804.08198 (2018).
9. Escolano, C., Costa-jussà, M., Fonollosa, J. Towards interlingual neural machine translation. arXiv: 1905.06831 (2019)
10. Arivazhagan, N., Bapna, A., Firat, O, Lepikhin, D., Johnson, M., Krikun, M., Chen, M., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., Wu, Y. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. arXiv:1907.05019v1 (2019).
11. Zeta Alpha Interview on Youtube: Angela Fan explains NLLB-200, <https://www.youtube.com/watch?v=IJZE7LikM3c> ; particularly around the 20-minute mark.
12. do Carmo, Félix. Debunking a few machine translation myths. CTS lecture uploaded on Youtube. <https://www.youtube.com/watch?v=Qap07fV5IgI> (2022).
13. Piantadosi, Steven T & Felix Hill. 2022. Meaning without reference in large language models. arXiv preprint arXiv:2208.02957
14. Melby, A. and Kurz, C. Data: Of Course! MT: Useful or Risky. *Translators: Here to Stay! MultiLingual* magazine (). <https://multilingual.com/issues/november-december-2021/data-of-course-mt-useful-or-risky-translators-here-to-stay/> (2021).
15. NLLB Team. No Language Left Behind: Scaling Human-Centered Machine Translation. <https://doi.org/10.48550/arXiv.2207.04672> (2022).
16. Bau, Anthony, et al. "Identifying and controlling important neurons in neural machine translation." arXiv preprint arXiv:1811.01157 (2018).
17. Macklovitch, E. The Translator's Workstation Project... in Plain Prose. In: *Proceedings of the 32nd Annual Conference of the American Translator's Association, Salt Lake City (1991)*.
18. Bawden, Rachel et al. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, Volume 1*, pages 1304-1313.

Automatic Detection of Omission in Comparative Literary Translation

Amal Haddad Haddad^[0000-1111-2222-3333]

¹ Universidad de Granada, C. Puentezuelas, 55 18002, Spain
amalhaddad@ugr.es

Abstract. Omission is considered a controversial issue in translation research. On the one hand, it is regarded as one of the common translation techniques used in cases of non-equivalence or implicit conveyance of meaning. On the other hand, it may be viewed as a sign of failure of the translator to render the Original Texts (OT) properly into the Target Languages (TL). Moreover, in some cases it may be considered as a parameter of manipulation and censorship. For this reason, when carrying out comparative translation research, the detection of omission and its analysis is one of the key elements to evaluate a translation, and to gain a full understanding of the translation decisions taken by a translator. In most cases, the process of detecting cases of omission in comparative research is carried out by manually annotating the Target Text (TT) in comparison to the OT, an arduous and time-consuming task, above all in long and extensive texts, such as some literary texts. For this reason, in this case-study, we use an alternative semi-automatic method to detect omission in translation research, and we use corpus analysis to provide results. Finally, we propose the creation of a new and more appropriate tool for the precise and automatic detection of omission, aimed at helping to obtain more results and a wider perspective in comparative literary translation studies.

Keywords: Computer Assisted literary Translation, Omission, Corpus Analysis.

1 Introduction

Omission is considered a controversial issue in translation research. On the one hand, it is considered as a solution in cases of non-equivalence or implicit conveyance of meaning (Baker, 2011), which has a cohesive function (Abdullatif, 2020). On the other hand, it is regarded as “a translation error where the translator fails to render a necessary element of information from the source text in the target text” (Delisle *et al.*, 1999: 165). Moreover, in some cases, it may be considered as a parameter of censorship (Klimovich, 2016). In some occasions, omission may be considered an accidental mistake and an unintentional error committed by the translator or by the editor (Melamed, 1996: 764). Baker and Saldanha (2009: 4) define omission as “the elimination or implicitation of part of the text”, while other researchers define it as “translation loss” (Dickins *et al.*, 2017) or “zero translation” where parts of the original texts are simply omitted or not included in the target text (Alrumayh, 2021: 1).

One of the most common approaches in translation studies is comparing the original texts with their translation and/or translations to comprehend which techniques have been used and the reason behind choosing them during the translation process. This approach is called close reading (Hayles, 2007; Youdale, 2019). One of the translation techniques which researchers try to focus on and detect in their studies is omission. When conducting comparative literary translation research, detecting omission is very important as it reveals the different preferences of different translators towards the Target Texts (TT) and in some cases, it gives information about their sociocultural and ideological tendencies (Klimovich, 2016).

Omission techniques are also relevant to censorship studies. This is due to the fact that readers can be manipulated by telling them only half-truths (Dimitriu, 2004: 174). Translators submissively avoid “anything that might shock the target audience or run against its shared beliefs” (Dimitriu, 2010: 174) to avoid clashes with dominating target cultural norms (Klimovich, 2015). Some researchers consider intentional omission as a direct strategy of censorship (Klimovich, 2016; Cámara, 2016; Alimen and Kalaycio, 2021) leading in some occasions to a shift in characterisation of the main characters, by trying to offer accessibility and acceptability of the target text in the target culture (Xiaoli, 2019: 204).

Baker and Saldanha (2009: 289) define censorship as “a coercive and forceful act that blocks, manipulates and controls cross-cultural interaction in various ways”. Other authors such as Leonardi (2008: 481) describe it as an “expression of ideology” in a sociocultural context and define it as: “a form of control over the readers which results in the manipulation or rewriting of the source text(s)”. According to Izwaini (2017: 47), the motive of censorship may be imposed by authorities like governments due to religious or sociocultural reasons, or may be practiced by translators themselves to fulfil with and respect the sociocultural value system, or the so called, self-censorship, defined as “an individual ethical struggle between self and context” (Santaemilia, 2008: 221). For this reason, Tymoczko (2000) considers omission as a practise of engagement in translation which implies an activist component.

The translation of Children and Youth Literature (CYL) is also subject to censorship and manipulation. The reason behind that is due to the fact that “both the target culture and society may decide what is wrong and what is acceptable for their children” (Leonardi, 2020: 26-27), and since this kind of literature tends to be orientated towards creating a particular image of childhood within the sociocultural contexts (Oittinen, 2006: 41). That is why CYL is considered an ideal field for censorship related studies (Giugliano and Hernández, 2019: 314), as they belong to both literary and educational field (Shavit, 1994: 11).

In the majority of research carried out until the moment to compare original texts with their translations, researchers do that by reading and manually annotating the original work and its translation and/or translations. This task is time consuming, labour-intensive and in many occasions could lead to the loss of relevant information due to inaccuracy, mainly in long and voluminous texts, such as some literary texts.

For this reason, in this paper, we propose using a combination between the close reading approach and the distant reading approach (Hayles, 2007; Moretti, 2013;

Youdale, 2019) which implies the use of new technologies, such as CAT tools and corpus analysis tools to acquire new insights into more comprehensive results in translation studies. With this objective, we suggest an automatic approach to detecting omissions in literary translated works with the help of the CAT tool Trados Studio 2021. This software was designed to help translators during the translation process by generating reusable Translation Memories (TM) that make translators benefit as much as possible from previous translations (Mitkov, 2022: 367); however, in this case-study, this programme is used to detect omissions techniques implemented by translators at complete segment level. Afterwards, we use sketch Engine as the main corpus analysis tool to further understand the manipulation techniques implemented by the translators by comparing two translations of the same OT.

1.1 Computer-Assisted Translation Tools and Literary Research

Traditionally, comparative literary studies have been carried out by manually comparing and annotating source texts and the equivalent translated texts in one or more languages. Despite the existence of a software which promises the automatic detection of omissions in translated texts, such as ADOMIT, which is an algorithm for automatic detection of omissions in translations (Melamed, 1996), very little research have been done to provide information on its use amongst translators and researchers and its efficiency in translation studies and research. Furthermore, this software is outdated and for the best of our knowledge, it is no longer available for its use.

Additionally, in spite of the fact that CAT tools and TMs revolutionised the work of translators in the last three decades (Mitkov, 2022: 364) and redefined translation competence (Zhang and Cai, 2015: 433), their capacity in providing a different perspective and in streamlining the analysis processes in literary translation and research has also been underestimated.

Previous research focused on the use of CAT tools to improve translation processes and above all, to enhance the workflow of translation tasks (Mitkov, 2022: 364). For instance, Youdale and Rothwell (2022) show how some translators use CAT tools to enhance the productivity and workflow of the process of translation of literary texts by using TMs. These authors state that in the last two years, there has been a slight shift in the attitude of some literary translators towards the use of CAT tools (Youdale and Rothwell, 2022: 383). Authors like Alcina (2008: 90) highlight that translation technologies help in making the translator's job easier, and at the same time, facilitate the research and teaching of translation activities. In a parallel way, other researchers such as Youdale (2019) underline the importance of using CAT tools and corpus analysis to understand new aspects in literary research. Other studies like Horenberg (2019) show the viability of using CAT tools in the phases of pre-analysis of source text as well as in its translation and retranslation; however, she also affirms that very little attempts have been carried out for the moment to find alternative ways to help in modernising literary translation tasks and research. For this reason, more research is still required to open new pathways in the new field of Computer-Assisted Literary Translation (CALT) (Youdale and Rothwell, 2022: 384) and to make the most out of the available resources and software. This type of studies may help in developing

more useful instruments and tools to help translators and researchers in the field of translation in the near future and may be useful to improve studies related to quality assessment in machine translation (Toral & Way, 2018; Mutal *et al.*, 2020) and in improving post-editing of literary texts (Moorkens *et al.*, 2018).

2 Materials and Methods

2.1 Materials

As a case-study, we carried out a comparative research on the epistolary novel, *Daddy-Long-Legs* and two of its translations into Arabic. *Daddy-Long Legs* was written originally by the American writer Jean Webster in 1912 and is still considered one of the symbols of the American national identity (Phillips, 1999: 79). This novel is still proving to be an international success until our current days as it is continuously being reedited and retranslated into different languages, and it has been adapted into stage and into screen. Since this novel is still relevant to CYL and popular in many societies and cultures, we consider its study and analysis as relevant in the field of translation and censorship.

Daddy-Long-Legs is classified as a youth literature novel (Guadamillas, 2019: 204; Hermida *et al.*, 2020: 10). It narrates the story of Jerusha Abot (Judy), a girl who was brought up at the orphanage of John Grier Home until she was 18. One of the trustees heard about her talent in writing and promised to finance her studies at college to become a writer, with the condition of receiving a monthly letter from her, describing her advances in her career and education. The trustee did not want to reveal his identity and said that he will never reply to her letters. On the day the trustee left the orphanage, Judy noticed only a glimpse of his shadow in the dark projected on a wall, and she starts calling him mockingly Daddy-Long-Legs, hence the title of the novel. When Judy started her life at college, she not only started writing one letter a month, but she used to send letters on weekly or daily basis describing all the details of her daily life. Through those letters, the educational, cultural, emotional, social and ideological growth of Judy is made tangible and visible. In the last two decades, different studies focused on the psychological and educational analysis of this novel from different points of views (Chang *et al.* 2010; Fitranti and Wedawati 2021; Guo, 2016) and also some studies analysed the figurative language that lies behind (Ramadhan, 2022). This shows that this novel is still relevant today.

In crosslinguistic and comparative research, previous studies have been carried out to compare *Daddy-Long-Legs* with its corresponding translations into different languages. For example, Sharifi and Karimnia (2014) analysed the translation of the translated book in comparison to the film dubbing in Persian language, by using the critical discourse analysis approach. Rahbar *et al.* (2013) identified the ideological content of the novel and study the dimension of censorship in the translations of the novel published in Iran, before and after the Islamic Revolution. Other authors such as Alimen and Kalaycioğlu (2021) compare two translations of the novel into Turkish adapted to children. However, for the best of our knowledge, no studies have been

carried out to analyse any Arabic translations of the novel and none of the studies implemented new technologies to compare results.

For this reason, in this case-study, we compare two translations of *Daddy-Long-Legs* into Arabic: the first translation entitled “أبي طويل الساقين” “*aby ṭawyl assāqayn*¹” [Daddy Long Legs²] was by Samir Mahfouz Bashir, published by the National Center for Translation in Egypt in 2009. The second translation, entitled “صاحب الظل الطويل” “*ṣāhib aẓẓill aṭṭawyl*” was by Buthaina Al-Ibrahim, published by Takween in Kuwait in 2018.

2.2 Methods

For the implementation of this case-study, first of all, the three versions of the novel were converted into an editable format, i.e. the original English and the two translations into Arabic. For this reason, we used the character recognition program I2PDF3 which allowed for the conversion of the scanned text into an editable format. The texts then were revised to guarantee the texts were correctly digitalised and legible and that no parts of the texts were lost in the process. Afterwards, the texts were imported into the CAT tool program Trados Studio 2021 with the objective of aligning the translations with the original text in English to create a parallel corpus. The automatic alignment offered by Trados Studio 2021 was revised by adjusting the segments and realigning them when needed. The alignment was applied at sentence and paragraph level. The split segment option was used when part of the sentence was omitted so that the unaligned segments would contain only the parts that were totally omitted. The untranslated segments were left unaligned.

When the two translations were adequately aligned, we used the option of identifying all null segments available in Trados Studio 2021. The functionality in Trados Studio 2021 appears in the alignment window as shown in Figure 1.

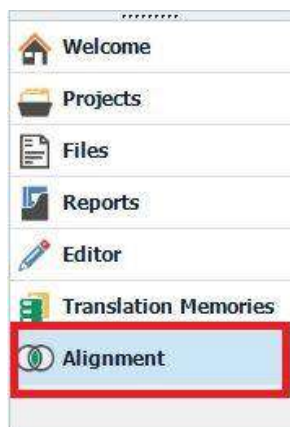


Fig 1. Alignment window in Trados Studio 2021

¹ Transliteration of Arabic text is provided between quotations.

² Literal translation of the Arabic text is provided between square brackets.

³ Available from: I2PDF: <https://www.i2pdf.com/es>

From that window, the functionality of *Select the alignment status, quality or connection type to go to* was selected as can be seen in Figure 2. Afterwards, the unconfirmed segments option was selected, as can be observed in Figure 3.



Fig 2. Select the alignment status, quality or connection type to go to

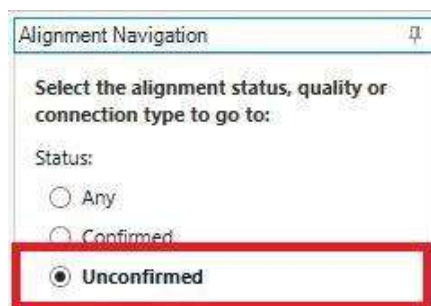


Fig 3. Select the unconfirmed segments option

This way all the segments that were not translated in any of the two translations into Arabic were identified by searching for null alignments in Trados 2021. By using this technique, it was possible to identify all the omissions of all null segments. Those omissions were verified in the original texts to guarantee that those segments were not omitted by mistake during the previous automatic processing. However, this technique is only helpful to detect the omission at the level of a whole segment and not on the level of subsegments.

For this reason, and in order to obtain more insightful and detailed comparative results, the aligned texts were inserted in the Sketch Engine corpus analysis tool (Kilgarriff, 2014)⁴ with the aim of comparing the OT and the two translations in Arabic. In other words, a bidirectional parallel corpus analysis was possible, using both a top down and bottom up strategy. The top down strategy refers to the previous close reading to unearth the meaning and comprehend the general settings and context of the novel in English and the two TTs. The bottom up strategy involves using corpus analysis tool, and functionalities such the *wordlist* and *keyword list* to detect any dif-

⁴ Available from: <http://www.sketchengine.eu>

ferences, either comparing the OT with its translations or comparing the lists in the two TTs.

Finally, the parallel concordances functionality in Sketch Engine was used to compare and analyse the techniques of translation used in different excerpts of the texts. This was carried out by comparing concordance in the two TTs. We compared how each translator rendered the segments that were omitted by the other translators to see whether there were evidences of manipulation or censorship. In other words, when an omission of a whole segment in T1 or T2 was observed, the parallel concordance option was used in Sketch Engine, comparing the OT with the other translation. This function helped in examining the omission at subsegment level.

In this phase, the study of Rahbar et al. (2013) in which the authors identified the ideological contents of the novel was also used to directly analysis manipulation techniques. For example, when Rahbar et al. (2013) detected omission due to ideological reasons in Persian, we compared those segments by means of the parallel concordances option in both T1 and T2. This helped in determining the manipulation dimension at word level and not only the omission techniques.

3 Analysis of Results

Firstly, the paragraphs and sentences which were omitted were identified by highlighting the null segments with the help of Trados Studio 2021. The result of this process indicated that the translation of Samir Mahfouz Bashir, to which we will refer as T1, has 62 instances of omissions, some of them are sentences and others are whole paragraphs. On the other hand, the translation of Buthaina Al-Ibrahim, to which we will refer to as T2 in this study, has no omissions at sentence and paragraph scales.

Secondly, and based on our cultural background and knowledge about the sociocultural context and norms in the Arab world, we classified the reasons of using omission techniques in T1 in five categories: a) omissions related to unacceptable social behaviour, above all, related to relations between men and women; b) omissions related to religious information; c) omissions related to ideological references; d) omissions related to unacceptable moral conduct; e) omissions due to linguistic reasons. Table 1 shows the frequency of omissions associated to each category.

Table 1. frequency of omission and motive of omission in T1

Motive of omission	Frequency
unacceptable social behavior	23
religious information	5
ideological references	4
unaccepted moral conduct	22
linguistic reasons	8
Total	62

After carrying out the analysis, it has been observed that social behaviour and moral conduct are the most frequent reasons of omissions in the translation, with a frequency of 23 and 22 occurrences respectively. In the following, examples on each motive will be given.

1. Jimmie McBride is going to teach me how to ride horseback and paddle a canoe, and how to shoot and--oh, lots of things I ought to know. It's the kind of nice, jolly, care-free time that I've never had; and I think every girl deserves it once in her life (p. 51).
2. I didn't know that people used to be monkeys and that the Garden of Eden was a beautiful myth (p. 14).
3. Seems a little early to commence entertaining, doesn't it? A friend of Pepys devised a very cunning manner whereby the king might pay his debts out of the sale to poor people of old decayed provisions. What do you, a reformer, think of that? I don't believe we're so bad today as the newspapers make out (p. 82).
4. Oh, you see, I know! You're a snappy old thing with a temper (p. 14).
5. I'd hate to retoumer chez John Grier (p. 41)

Example (1) shows a case of omission of a whole paragraph, as it describes a situation in which a man and a woman would have a close relationship doing certain activities together, which are not accepted in some cultures. In this case, Jimmie McBride is one of Judy's friends and the brother of her best friend Sally, and he invited Judy to spend the summer with him in to teach her certain activities. These types of omissions were frequent also in scenes where Judy and Jervis Pendleton were together.

Example (2) shows a case of omission in order to hide information that is considered contradictory to religious teachings. In this particular example, the omission was implemented to avoid telling information relevant to stating that the origin of people is monkeys, and questioning the veracity of existence of the Garden of Eden.

In the case of Example (3), it shows omission due to ideological contents. In this example, the word "reformer" is the clue. The translator estimated that he shouldn't include this type of information related to a particular political trend.

With respect to Example (4), it shows a disrespectful behaviour from part of the main character towards the trustee, who she supposes he is an old man and the way she picked her words is considered disrespectful.

Finally, in cases like Example (5) the linguistic difficulty which led to the use of omission was due to mixing English with French which the translator preferred to omit due to the additional difficulty of rendering this information in Arabic.

After detecting and analysing all the cases of omissions in T1, we compared the T1 and T2 by using the *parallel concordances* option available in Sketch Engine. This was used above all to analyse the techniques used by the translator in T2 to render the parts of the novel in which there were omissions in T1. On the other hand, by using this method, we also verified how the two translators rendered the instances where Rahbar *et al.* (2013) identified ideological content. Table 2 shows some of the words

or sentences that went through alteration of meaning techniques, comparing the original text in English with its equivalents in T1 and T2.

Table 2. Manipulated words or sentences in T1 and T2 in comparison to the OT

Original text	T1	T2
<i>cotillion</i>	الرقص <i>arraḡṡ</i> [the dance]	رقصة الكوتليون <i>raḡṡat alqwtlywn</i> [the cotillion dance]
<i>They are pigs!</i>	-	إنها قذرة! <i>Innaha qadhira</i> [It's dirty]
<i>anarchists</i>	المحافظين <i>almuḡāfizyn</i> [The conservatives]	الفوضويين <i>alfawḡawyīn</i> [The anarchists]
<i>plutocrat</i>	رأسمالية <i>Ra'smālyā</i> [Capitalism]	بلوتوقراطية <i>blūtwḡratya</i> [Plutocracy]
<i>Yours ever</i>	المطيعه دائما <i>almuḡy'a da'iman</i> [The always obedient]	لك أبدا المخلصه <i>almukhliṡa laka abadan</i> [the always faithful]

Table 2 shows some of the examples where the techniques of meaning alteration were used either in T1 or in T2. In the case of the word *cotillion*, which is the name of the dance which implies close interaction between a man and a woman, it was mentioned by Judy as it was the dance that took place during a party. In T1, the translator used the metonymy “الرقص” “*arraḡṡ*” [the dance] as a way of euphemism, instead of mentioning the name of the dance as can be seen in T2.

In the case of the omitted sentence in T1 “They are pigs!”, the translator in T2 decided to translate it as “إنها قذرة” “*Innaha qadhira*” which means “it’s dirty”. The omission in T1 indicated that the translator considered mentioning pigs as inappropriate in the TT, while the translator in T2 added the connotation of pigs being dirty.

With respect to the word “anarchist” which appeared in the following context: “You know, I think I’ll be a Socialist, too. You wouldn’t mind, would you, Daddy? They’re quite different from Anarchists; they don’t believe in blowing people up” (p. 67), it is observed that the translator in T1 decided to change the meaning of the word into “المحافظين” “*almuḡāfizyn*” which means “the conservatives” while the translator in T2 used the word “الفوضويين” “*alfawḡawyīn*” which a similar equivalent of the original text. In the same way, the translator in T1, substituted the word “plutocrat” by the word “رأسمالية” “*Ra'smālyā*” which means “capitalism”, while in T2, the translator used the literal translation “بلوتوقراطية” “*blūtwḡratya*”. All those words are related to political ideology.

Finally, in T1, the translator changed the way in which the main character, Judy, finishes her letters, by substituting the sentence “yours ever” with “المطبعة دائما” “almuty‘a da’iman” which means “the always obedient”, while the translator in T2 used a more literal translation.

As shown in the previous examples, both translations of *Daddy-Long-Legs* went through omission or manipulation techniques. However, T1 contains obvious examples on omission and alteration of meaning due to ideological reasons. In this case, the use of the CAT tool Trados Studio 2021 helped in the automatic detection of omission at segment level and facilitated in combination with the Sketch Engine tool in identifying clear evidence of censorship and manipulation.

4 Conclusions

In this case-study, we aim at contributing to the new field of Computer-Assisted Literary Translation (Youdale and Rothwell, 2022: 384). For this reason, on the one hand, an automatic methodology using available CAT tools to detect omissions at segment level in literary texts is suggested. On the other hand, the use of corpus analysis methodology to analyse further translation techniques, such as alteration of meaning is also proposed. Those translation techniques are considered especially relevant to censorship studies and to literary translation in general.

Omission has a multifaceted nature and detecting it is very helpful in comparative translation studies; however, due to the lack of proper automatic or semi-automatic tool, most of the researchers in the literary field use manual analysis and manual annotation of original texts in comparison to their translations. This manual process is time-consuming and sometimes may lead to the loss of relevant information. For this reason, in line with the suggestions of Moretti (2013) and Youdale (2019), we applaud the idea of combining close reading and distant reading in literary translation research in order to speed up the process of analysis and obtain a wider perspective of more precise results and better informed translations.

With this objective, and due to the lack of proper tools to automatically align and detect cases of omission, we used of the CAT tool Trados Studio 2021, as we considered it as an alternative method to allow for the semi-automatic detection of omission after the proper alignment of the original text with its corresponding translation and/or translations.

As a complementary method, we also used corpus analysis tools such as Sketch Engine to compare word lists and its frequency as well as using the parallel concordance function to see how certain words or sentences were used in the translation, object of study.

As part of a case-study, we applied the methodology of both close reading and distant reading to study the translation techniques used in two Arabic translations of the epistolary novel, *Daddy-Long-Legs* written by Jean Webster in 1912. We focused above all on the techniques of omission and alteration of meaning. For this reason, we used the CAT tool Trados Studio 2021 for the semi-automatic detection of omissions and then we used the corpus analysis tool Sketch Engine to detect possible manipula-

tions of meaning. As a result, we detected 62 omissions in one of the translations into Arabic and we analysed the motivation behind the use of this technique. On the other hand, we detected some intentional manipulation of meaning in both translations in some of the sentences and words due to ideological motives. As a result, we conclude that the translation carried out by Samir Mahfouz Bashir is subject to censorship and alteration of meaning for ideological, religious and sociocultural reasons; however, we did not analyse whether it is due to self-censorship or for being imposed by other authorities, as this analysis is beyond the scope of this study. The use of those methods in combination with the close reading approach proved to give more insights and wider perspectives towards OTs and TTs.

On the other hand, we believe that in spite of the revolutionary role of new technologies such as CAT tools in performing translation tasks and in carrying out translation studies, these tools are far from perfect and suffer a set of shortcomings (Mitkov, 2022). Notwithstanding that Trados Studio 2021 helped in the automatic detection of omissions in this study, and it saved time, we still consider this method as insufficient as the process of aligning the two texts and making sure all the segments are correctly associated with its correct equivalent parts is also a time-consuming task and may be also subject to human error during the process of realignment of segments or due to the inappropriate alignment of segments. For this reason, we believe there is still a need to design more efficient software with the help of more sophisticated artificial intelligence programs to align texts automatically with their translations with major perfection, and have the functionality of detecting omissions in a more straight forward way, and not only at complete segment level, but also, at subsegment and word level. Having such a reliable instrument would not only help researchers in detecting omission as a translation techniques or as evidence of censorship, but it would also help translators themselves to make sure they have not committed any omission errors due to negligence. We also believe that studies resulting from applying this type of methodology would also give more insights that may help in improving quality related to omission approaches in machine translation (Mutal, *et al.*, 2020) and would also help in improving the quality of neural machine translation for literary texts (Toral & Way, 2018).

Finally, it is also important to highlight that the translation memories generated after the alignment process are also valuable. Those TMs would be useful in future translations of the same work and in retranslation of similar works of the same author. In the case of Daddy-Long-Legs, those TMs would help for example in translating *Dear Enemy*, which is the novel sequel to *Daddy-Long-Legs*. Those TMs would not only help in improving consistency of vocabulary, but may also help in maintaining the same style.

Acknowledgements

Funding was provided by an FPU grant (FPU18/05327) given by the Spanish Ministry of Education.

References

1. Abdullatif, M.: Omission as a Problem and a Solution in Literary Translation: A Study of Hemingway's Hills Like White Elephants and Joyce's The Sisters and Their Arabic Translations. *Journal of Languages and Translation*, 7(1), 16-41 (2020).
2. Alcina, A.: Translation Technologies: scope, tools and resources. *Target: International Journal on Translation Studies*, 20 (1), pp. 79-102 (2008).
3. Alimen, N. and Kalaycioğlu, E.: An Analysis of the Turkish translations of Jean Webster's Daddy-Long-Legs from the perspective of systemic affiliation. *TransLogos*, 4(2), pp. 70 – 97 (2021).
4. Alrumayh, A. Translation by Omission and Translation by Addition in English-Arabic Translation with Reference to Consumer-oriented Texts. *International Journal of Comparative Literature & Translation Studies* (1), pp. 1-10 (2021).
5. Baker, M.: *In other words: A Coursebook on translation*. Routledge (2011).
6. Baker, M. and Saldanha, G.: *Routledge Encyclopedia of translation studies*. 2nd edn. Routledge (2009).
7. Cámara Aguilera, E.: Traducción y asimetría: "Manolito Gafotas" y su traducción al inglés como ejemplo de intervencionismo. *Anuario de investigación en literatura infantil y juvenil: ANILIJ*, 14, pp. 23-42 (2016).
8. Chang, J., Masters, A., Avery, A. Werren, J. H.: A divergent Cardinium found in Daddy Long-Legs (Arachnida: Opiliones), *Journal of Invertebrate Pathology*, 105, pp-220-227, (2010).
9. Delisle, J., Lee-Jahnke, H. and Cormier, M. C.: *Terminologie de la traduction*. John Benjamins: Amsterdam & Philadelphia (1999).
10. Dickins, J., Hervey, S. and Higgins, I.: *Thinking Arabic translation*. 2nd edn. Routledge (2017).
11. Dimitriu, R.: Omission in translation, *Perspectives: Studies in Translatology*. *Studies in Translation Theory and Practice* 12 (3), pp. 163-175 (2004).
12. Fitranti, F. H., and Wedawati, M. T.: Self-concealment in Jean Webster's Daddy Long Legs. *Elite : English and Literature Journal*, 8(1), 1-14 (2021).
13. Giugliano, M. and Hernández Socas, E.: Ambivalence, Gender, and Censorship in two Spanish Translations of Little Women. *Meta*, 64(2), pp. 312–333 (2019).
14. Guadamillas Gómez, M. V.: Inolvidables. Grandes autoras que escriben para los pequeños. *AILIJ: Anuario de Investigación en Literatura Infantil y Juvenil*, 17, pp. 203-208 (2019).
15. Guo Y.: An Analysis of Judy's Personality of Independence in Daddy-Long-Legs, 11 pp. 3-4, (2016).
16. Hayles, K. N.: *Hyper and Deep Attention: The Generational Divide in Cognitive Modes*. Profession, pp. 187-199 (2007).
17. Hermida, C., Couso, L. B., Bayerque, M. A.: Jóvenes y literatura: cruces entre el campo editorial y escolar. In proceedings of 2nd Congreso Latinoamericano de Comunicación de la UNVM. Instituto Académico Pedagógico de Ciencias Sociales. Universidad Nacional Villa María, Villa María, Córdoba (2020).
18. Horenberg, L. W.: *Using CAT in Literary Translation - How Tools May Support Translators in Source-Text Analysis, Translation and Retranslation: A Case Study of "Mr Love-day's Little Outing"*. MA thesis (2019).
19. Izwaini, S.: Censorship and manipulation of subtitling in the Arab world. In Jorge Díaz Cintas y Kristijan Nikolić, editors-in-chief, *Fast-Forwarding with Audiovisual Translation*. *Multilingual Matters*, pp. 47-57 (2017).

20. Kilgarriff, A., Baisa, V., Bušta, J., Jakubiček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. *Lexicography*, 1: 7-36. (2014).
21. Klimovich, N.: Manipulation in Translation (Exemplified by the Intertextual Elements' Translation). *Journal of Siberian Federal University. Humanities & Social Sciences* 2 (8), pp. 244-251 (2015).
22. Klimovich, N.: Manipulative Strategies in the Translations of Literary Texts Carried Out in the Soviet Union. *Journal of Siberian Federal University. Humanities & Social Sciences*, 3 (9) pp. 543-550 (2016).
23. Leonardi, V.: Power and control in translation: between ideology and censorship. *Language, Communication and Social Environment*, 6, pp. 80-89 (2008).
24. Leonardi, V.: Ideological manipulation of children's literature through translation and re-writing: travelling across times and places. Palgrave Macmillan (2020).
25. Melamed, Dan, I.: Automatic Detection of Omissions in Translations. In COLING 1996 (2): The 16th International Conference on Computational Linguistics, pp. 764–769 (1996).
26. Mitkov, R.: Translation Memory. In Sharon Deane-Cox and Anneleen Spiessens, editors-in-chief, *The Routledge Handbook of Translation and Memory*. Routledge: Basingstoke (2022).
27. Moorkens, J., Toral, A., Castilho, S., Way, A.: Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7(2), pp. 240-262 (2018).
28. Moretti, F.: *Distant Reading*. London : Verso (2013).
29. Mutal, J., Gerlach, J., Bouillon, P., and Spechbach, H.: Ellipsis translation for a medical speech to speech translation system. 22nd Annual Conference of the European Association for Machine Translation (EAMT) (2020).
30. Oittinen R. No Innocent Act: On the Ethics of Translating for Children. In Van Coillie J, Verschueren WP, editors, *Children's Literature in Translation. Challenges and Strategies*. Manchester, UK & Kinderhook, USA. p. 35-46 (2006).
31. Philips, Anne K.: "'Yours most loquaciously': Voice in Jean Webster's *Daddy-Long-Legs*". *Children's Literature* n° 27, pp. 64-86 (1999).
32. Rahbar, M., Bateni, B., and Abad-Najaf, R.: Ideological manipulation in translation: A case study of Jean Webster's "*Daddy Long Legs*". *International Journal of Language Learning and Applied Linguistics World (IJLLALW)* 4(4), pp. 373-381 (2013).
33. Ramadhan, D. R.: Analysis of figurative language translation in Jean Webster's novel *Daddy-Long-Legs*. Diploma thesis, Universitas Nasional (2022).
34. Santaemilia, J.: "The Translation of Sex-Related Language: The Danger(s) of Self-Censorship(s)". *TTR* 21, no. 2: 221–252. (2008).
35. Sharifi, L. and Karimnia, A.: Differences between Webster's *Daddy-Long-Legs* Translation for Publication and Animation: Insight From Van Dijk's Ideology Framework of Critical Discourse Analysis. *Modern Journal of Language Teaching Methods (MJLTM)* 4(4), pp. 163-169 (2014).
36. Shavit, Z.: Beyond the Restrictive Frameworks of the Past: Semiotics of Children's Literature — A New Perspective for the Study of the Field. In Hans-Heino Ewers, Gertrud Lehnert, Emer O'Sullivan, editors-in-chief, *Kinderliteratur im interkulturellen Prozeß*, pp. 1-6 (1994).
37. Toral, A., Way, A.: What Level of Quality Can Neural Machine Translation Attain on Literary Text? Springer International Publishing, pp. 263-287 (2018).
38. Webster, J.: *Daddy-Long-Leg*. Project Gutenberg: Rainfield (2003).
39. Xiaoli, W.: Omission and its Impact on Character Reshaping in Literary Translation: A Case Study of *Wolf Totem*. *Interlitteraria* (2019).

40. Youdale, R.: *Using Computers in the Translation of Literary Style: Challenges and Opportunities*. 1st edn. Routledge (2019).
41. Youdale, R. and Rothwell, A.: Computer-assisted translation (CAT) tools, translation memory, and literary translation. In Sharon Deane-Cox and Anneleen Spiessens, editors in chief, *The Routledge Handbook of Translation and Memory*. Routledge, London (2022).
42. Zhang, C. and Cai H.: On Technological Turn of Translation Studies: Evidences and Influences. *Journal of Language Teaching and Research*, 6 (2), pp. 429-434 (2015).

Comparing Interface Designs to Improve RSI platforms: Insights from an Experimental Study

Muhammad Ahmed Saeed^[0000-0001-7452-5836], Eloy Rodríguez González^[0000-0003-4045-0332],
Tomasz Korybski^[0000-0003-2353-0816], Elena Davitti^[0000-0002-7156-9275] and Sabine
Braun^[0000-0002-6187-3812]

University of Surrey, Guildford, Surrey GU2 7XH, UK
m.a.saeed@surrey.ac.uk

Abstract. Remote Simultaneous Interpreting (RSI) platforms enable interpreters to provide their services remotely and work from various locations. However, research shows that interpreters perceive interpreting via RSI platforms to be more challenging than on-site interpreting in terms of performance and working conditions [1]. While poor audio quality is a major concern for RSI [2,3], another issue that has been frequently highlighted is the impact of the interpreter's visual environment on various aspects of RSI. However, this aspect has received little attention in research. The study reported in this article investigates how various visual aids and methods of presenting visual information can aid interpreters and improve their user experience (UX). The study used an experimental design and tested 29 professional conference interpreters on different visual interface options, as well as eliciting their work habits, perceptions and working environments. The findings reveal a notable increase in the frequency of RSI since the beginning of the COVID-19 pandemic. Despite this increase, most participants still preferred on-site work. The predominant platform for RSI among the interpreters sampled was Zoom, which has a minimalist interface that contrasts with interpreter preferences for maximalist, information-rich bespoke RSI interfaces. Overall, the study contributes to supporting the visual needs of interpreters in RSI.

Keywords: Remote Simultaneous Interpreting (RSI), User Experience Questionnaire (UEQ), User Experience (UX)

1 Introduction

In recent decades, technology has intersected with interpreting in various ways, notably in different modalities of distance interpreting and computer-assisted interpreting [2]. In relation to distance interpreting, the uptake prior to the COVID-19 pandemic has been uneven across different interpreting settings. In public-service settings, there has been an increase and diversification of distance interpreting, which further surged during the pandemic [4,5]. However, in conference interpreting, the adoption of distance interpreting was limited before the pandemic despite remarkable technological advancements in remote simultaneous interpreting (RSI). The limited adoption was

attributed to several factors such as poor audio quality and reluctance among conference interpreters to use new, cloud-based communication technologies [2,6].

Initially, RSI was conducted from traditional booth-based environments, such as interpreting hubs. These hubs were connected to remote clients and provided interpreters with the conventional conference interpreting equipment, including the interpreter's hardware console [6]. By contrast, the new generation of cloud-based simultaneous interpreting delivery platforms which emerged in the years before the pandemic offered interpreters a virtual console and a work environment that simulated traditional interpreting booths [6]. While the new platforms could still be used from within a traditional interpreting booth, they also had the potential to enhance the flexibility of RSI. At the same time, cloud-based RSI also raises several issues. A recent survey [1] indicates that interpreters believe that their performance and working conditions are worse under RSI conditions, corroborating previous findings that interpreters perceive RSI as more challenging than on-site interpreting [2]. While audio quality is highlighted by interpreters as a major factor affecting RSI [2,3,7], another issue that has been consistently highlighted in relation to both booth-based and platform-based RSI is the interpreters' visual environment [3].

To address this issue, the study reported here, which is part of a larger study focusing on various aspects of RSI, aims to explore the extent to which different aspects of an RSI interface, especially different visual aids and different approaches to presenting visual information, can support interpreters and enhance their user experience (UX). The study drew on an experimental design and tested a cohort of 29 professional conference interpreters in working with different visual interface options, examining the impact of three independent variables on the interpreter's UX:

- **Type of interface**, with a minimal interface design prioritizing the speaker and featuring hidden controls inspired by Zoom, and a maximal interface design displaying all functions and settings, inspired by bespoke RSI platforms.
- **Interpreter's view of the speaker**, with a close-up view option displaying only the speaker's face and a gesture view option showing the speaker's upper body and hand gestures in addition to their face.
- **Automatic Speech Recognition (ASR)**, with an ASR panel embedded within an RSI interface compared to an interface without ASR.

These visual variables were chosen because they have often been overlooked in RSI research in favour of investigating sound quality effects [2,3,4]. However, previous studies have highlighted the impact of nonverbal visual information, such as hand gestures, lip movement, and body language in interpreting [8]. Other research has emphasized the importance of addressing the interpreters' sense of presence by thoroughly analyzing their visual requirements [3]. Furthermore, studies on the use integration of ASR in the interpreting workflow have underscored the positive impact that ASR may have for interpreters [9]. In addition to exploring the interpreters' preferences regarding soft visual aids, we also examined how their attitudes towards RSI were influenced by demographic factors and by their experience in both simultaneous interpretation (SI) and RSI.

This article investigates interpreters' preferences and usage patterns of various RSI platforms. We also examine the relationship between participants' most frequently used platforms and their UX, as measured by the User Experience Questionnaire (UEQ) [10] ratings for the tested interface design philosophies, specifically minimalist and maximalist approaches. For brevity, we will not discuss the speaker view and ASR study variables in this article.

2 Methodology

The experimental study reported here used a within-subject design. It was preceded by a focus group with interpreters to inform its design, and supplemented by follow-up interviews with selected interpreters, but the findings from the focus group and interviews are not presented here. For the experimental study, which was conducted between February 2022 and May 2022, we recruited professional interpreters who had at least 400 hours of SI experience and at least 20 hours of RSI experience. The study was conducted online using Qualtrics [11] to guide the participants through the experiment, presenting them with the interpreting tasks, source speeches and questionnaires. In addition, Zoom was used to record the participants' interpretations. The experiment could be performed within a 24-hour window, according to participants' availability.

In the part of the study that related to soft visual aids for the interpreter, a two-by-two factorial design was used, involving the simultaneous manipulation of two independent variables. The interpreters were requested to interpret a 25-minute speech, which was divided into two segments. Each segment featured one of two interface designs (minimal/maximal), which were counterbalanced among the interpreters, i.e., some interpreters were initially presented with the minimal design, while others began with the maximal design. Additionally, within each segment, the speaker view (close-up/gesture view) was switched halfway through to ensure that each interpreter experienced both speaker views within each interface design. The order of the different speaker views was also counterbalanced to minimize potential order effects [12]. A publicly available recording of a real-life speech was selected. A crucial selection criterion was that it would display the speaker in the video frame, enabling us to create a contrast between a close-up view and a gesture view.

A pre-experiment questionnaire focused on the interpreters' RSI-related perceptions, preferences, and work setups. A series of short 'in-experiment' questionnaires, administered after each interpreting task, elicited UX ratings based on the short version of the UEQ [13] and qualitative comments from the participants regarding the interfaces presented. Finally, a post-experiment questionnaire elicited the interpreters' subjective preferences towards the speaker view part of the experiment and any suggestions they had for improving it. We employed a mock RSI interface with two versions, maximal and minimal (see Figure 1) because real-life simultaneous interpreting platforms do not offer users the ability to create multiple distinct versions of the interface as required for our study. For further information on the interface design process, readers can refer to [14].



Fig. 1. Maximal interface design (left) and minimal interface design (right)

3 Characteristics of the sample

Out of a total of 35 professional interpreters recruited for the study, 29 successfully completed all of the tasks. Of these 29 participants, 19 identified as female and 10 as male. The majority of participants (89.6%) fell in the range of 30-59, with 8 participants aged 30-39, a further 8 participants aged 40-49, and 10 participants aged 50-59. Two participants were under 30 and one was over 60. Our study's age and gender distribution shows similarity to other recent studies [1,7], providing contextual information for the current study and facilitating comparison with similar investigations.

3.1 Interpreters' work experience and preferences

Of the 29 study participants, 23 had more than 1,200 hours of SI experience, while 6 had between 400 and 1,200 hours of SI experience. In terms of RSI experience, 27 participants had over 50 hours, and 2 had between 20-50 hours of RSI experience. The participants interpreted from English into various languages, including Spanish, French, German, Italian, Polish, Latvian, and Cantonese.

Before the pandemic, most interpreters in our sample primarily worked on-site. However, the majority of interpreters in our sample reported an increase in their remote work frequency after the beginning of the pandemic. These findings corroborate other recent findings [1] and suggest that the pandemic has had a significant effect on the work patterns and experiences of interpreters, leading many to rely increasingly on remote work.

The majority of participants in our study preferred on-site work to remote work. Of our 29 participants, 18 either strongly or slightly preferred onsite work, while 7 either strongly or slightly preferred remote work, and 4 had no preference. However, our findings also show that preferences varied to some extent by gender and age. Female participants were twice as likely as male participants to prefer RSI. This may be because RSI, by eliminating the need to travel, makes it easier for women to balance work and family commitments, among other possible factors [15].

Nevertheless, both genders considered RSI to be more difficult and hampering teamwork. In terms of age range, interpreters in the 50-59 age range expressed the

strongest preference for on-site work, despite generally having more experience with RSI than younger interpreters. An interesting contrast was, however, observed for this group. Although 90% of interpreters in this group preferred on-site work, over 50% of their actual workload consisted of RSI.

According to the participants' qualitative feedback, the major factors contributing to the preference for onsite work included better working conditions, the ability to physically interact with the speaker, and the opportunity to collaborate with colleagues face-to-face. Participants appreciated the opportunity to work in a physical workspace with a well-equipped infrastructure that provided them with an environment conducive to productivity. Conversely, the main reason cited for a preference for remote work was to avoid commuting, which was a significant source of stress and time consumption for participants.

3.2 Interpreters' platform experience and preferences

As depicted in Figure 2, our pre-experiment questionnaire results show that 28 out of 29 participants had used Zoom with the interpretation function, with 17 indicating it as their most frequently used platform. While 17 participants had used Interactio and Interprefy, only four and two participants, respectively, identified them as their most frequently used platforms. Our findings are consistent with other recent studies [1,7], suggesting widespread use of Zoom with interpreting function for RSI.

All participants also mentioned either having tested or having used more than one platform for their interpreting assignments, allowing the research team to gather valuable insights into comparative user experiences, preferences, and potential strengths and weaknesses of different RSI platforms.

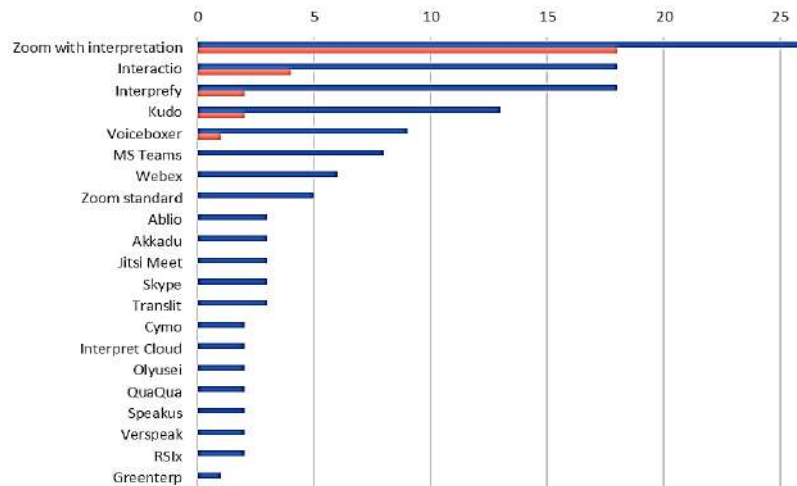


Fig. 2. Interpreters' platform experience: Blue=number of participants who reported using the platform; orange=number of participants reporting it as their most frequently used platform

The preferred platforms in our sample were Zoom with interpretation function and Interprefy. Both were mentioned as preferred platforms by 11 participants. Interestingly, while participants use Zoom more frequently than bespoke RSI platforms, they almost always prefer the functionality offered by bespoke platforms. They use Zoom mainly because it is imposed by event organizers due to cost or privacy concerns. Some interpreters in our sample appreciated its simplicity and stability, arguing that it has the basic features required for interpreting despite missing some functionalities.

Overall, the findings from our pre-experiment questionnaire suggest that the interpreters in our sample have a diverse range of experiences and preferences in relation to RSI and the platforms used for it. We consider this diversity to be a strength of our study as it ensures that the UX testing of visual aspects of an RSI interface is grounded in a variety of experiences and preferences.

4 Interface type: UEQ findings

The User Experience Questionnaire (UEQ) is widely used to evaluate UX across six dimensions, including attractiveness, efficiency, and novelty, and two meta-dimensions: pragmatic and hedonic quality, measuring perceived practical usefulness and user enjoyment respectively, using 26 item pairs rated on a seven-point Likert scale [9]. The UEQ is commonly used to evaluate prototypes in various domains. A short version of the UEQ, consisting of only eight item pairs and focusing on the two meta-dimensions, was developed for studies requiring quick completion and studies such as the present one, where multiple product or prototype versions need evaluation in a single session [10]. We selected the short UEQ for our study and expanded it by including three additional item pairs to measure specific dimensions of the user experience (UX) of RSI. These dimensions are Technical/Human, which evaluates the comprehensibility of the interface design from a human perspective, Alienating/Connective, which assesses whether the interface evokes any sense of alienation or whether it fosters a seamless workflow, and Unprofessional/Professional, which gauges the resemblance of the interface to conventional interpreting tools and RSI platforms.

4.1 Individual UEQ item pair scores

The individual UEQ scores for the two interface types (Figure 3) show that the feature-rich maximal interface outperformed the minimal interface in 9 out of 11 item pairs, particularly in Technical/Human, Boring/Exciting, and Usual/Leading edge. This suggests that users find the maximal interface more human-like, exciting, and cutting-edge. However, both interfaces scored similarly in Complicated/Easy and Confusing/Clear, indicating that participants appreciated the simplicity and clarity of the minimal interface despite their overall preference for the maximal interface.

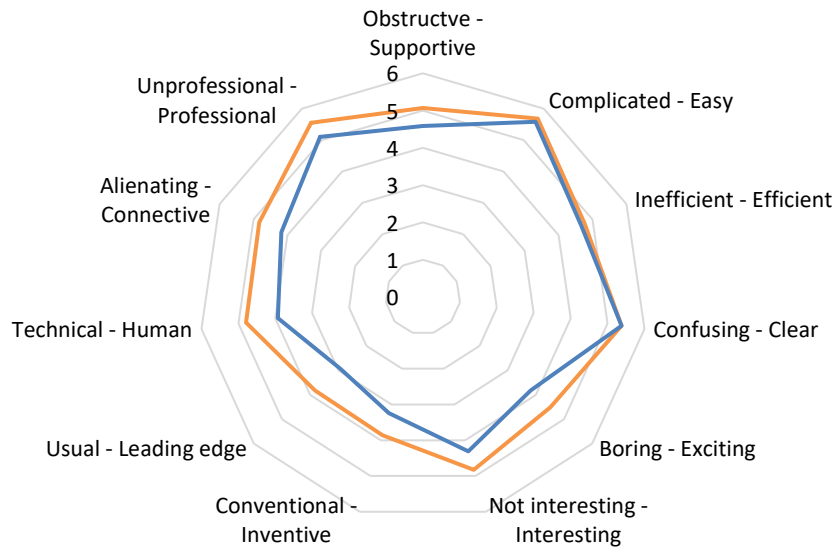


Fig. 3. UEQ item pairs; blue=minimal, orange=maximal interface

4.2 Overall UEQ scores

From the overall user ratings (Figure 4), it is clear that the maximal interface outperformed the minimal interface in terms of both usability and joy of use. Specifically, the maximal interface scored almost twice as high as the minimal interface in hedonic quality, which represents user enjoyment, while the minimal interface had a negative score for this meta-dimension. However, the maximal interface only had a slightly better score than the minimal interface in pragmatic quality, which pertains to the perceived practical usefulness of the interface. This suggests that while users find the maximal interface more enjoyable, both interfaces allow them to accomplish tasks effectively.

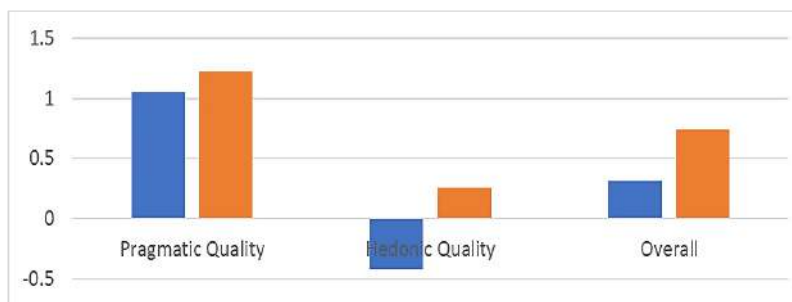


Fig. 4. Overall UEQ Scores; blue=minimal, orange=maximal interface

5 Discussion

This article raises several interesting discussion points. First, there is a notable contrast in the 50-59 age group, where 90% of interpreters preferred onsite work, yet over 50% of their actual workload consisted of RSI. This may suggest that interpreters in this age group perceive their working conditions as particularly adverse. It is possible that interpreters in this age group took on RSI work during the pandemic for financial reasons, i.e., due to a lack of onsite assignments during this time, but also due to increased concerns about travelling and potential exposure to COVID-19, even after the lockdowns, compared to younger interpreters.

Another point that emerges from the experience profiles of participants in all age groups is their preference for onsite work over remote work, stemming from better working conditions, the opportunity to physically interact with speakers, and face-to-face collaboration with colleagues. However, this preference comes with a trade-off in terms of increased stress and time spent travelling, which ultimately results in fewer assignments completed in the same amount of time. This raises the question of whether and under what circumstances a point could be reached where interpreters would consider both SI and RSI to be more or less equivalent.

Our exploration of visual aspects of the RSI interface suggests that visual aspects play a role in the interpreters' preferences regarding SI vs. RSI. Our comparison of interpreters' actual platform experiences and the User Experience Questionnaire (UEQ) ratings for each interface reveal a discrepancy between the minimalist interfaces (such as Zoom) that interpreters regularly use and the more comprehensive, bespoke RSI interfaces that they rate higher. Interpreters appear to prefer feature-rich and specialized RSI interfaces while often using more minimalist platforms in their daily work due to factors like cost, accessibility, and market penetration [1,7]. This discrepancy could be one of the factors contributing to the continued dislike of RSI, even though platform-based RSI offers more flexibility than traditional booth-based RSI, such as the ability to work from home.

6 Conclusion

In conclusion, the primary objective of our study is to explore user preferences and experiences related to different interface options, specifically focusing on improving RSI interfaces. We aim to investigate how visual aspects of an RSI interface contribute to the user experience of interpreters and to identify effective ways to display visual information that can enhance the overall UX. In this article, we first examined the preferences and usage patterns of 29 interpreters regarding various RSI platforms. Second, we analyzed the impact of different visual interface versions, a minimalist and a maximalist version, on the interpreters' user experience, drawing on the short version of the UEQ. The sample characteristics highlight the significant impact of the COVID-19 pandemic on interpreters' work patterns and experiences, leading to an increased reliance on RSI. However, the majority of interpreters still prefer on-site work, citing better working conditions, the ability to physically interact with speakers,

and the opportunity to collaborate face-to-face with colleagues as key factors contributing to this preference. The UEQ results indicate a preference for feature-rich and bespoke RSI interfaces when objectively evaluated, which contrasts with current practice where interpreters commonly use minimalist interfaces. The maximal interface outperformed the minimal interface in terms of usability and hedonic quality, suggesting that interpreters may benefit from interfaces that prioritize functionality and enjoyment. Our future publications will focus on the other two study variables, namely the speaker view and the inclusion of a source speech transcript created through ASR, as well as on the interview responses gathered from the participants after the experiment.

References

1. Buján M, Collard C. Remote Simultaneous Interpreting and COVID-19: Conference Interpreters' Perspective. In: Liu K, Cheung AKF, editors. *Translation and Interpreting in the Age of COVID-19. Corpora and Intercultural Studies*, vol. 9. Singapore: Springer; 2022. p. 93-108.
2. Braun S. Technology and interpreting. In: O'Hagan M, editor. *The Routledge Handbook of Translation and Technology*. London: Routledge; 2019. p. 271-287.
3. Ziegler K, Gigliobianco S. Present? Remote? Remotely present! New technological approaches to remote simultaneous conference interpreting. In: Fantinuoli C, editor. *Interpreting and technology*. Berlin: Language Science Press; 2018. p. 119-139.
4. Braun S, Al-Sharou K, Temizöz Ö. Technology use in language-discordant interpersonal healthcare communication. In: Gavioli L, Wadensjö C, editors. *The Routledge Handbook of Public Service Interpreting*. London: Routledge; 2023. p. 159-178.
5. Singureanu D, Braun S, Hieke D, Gough D. *Managing the demands of video-mediated court interpreting: strategies and the role of emotional intelligence [PhD thesis]*. Surrey (UK): University of Surrey; 2023.
6. Seeber KG, Keller L, Amos R, Hengl S. Expectations vs. experience: A survey of interpreters working with videoconferencing. *Interpreting*. 2019;21(2):270-304.
7. Chmiel A, Spinolo N. Testing the Impact of Remote Interpreting Settings on Interpreter Experience and Performance. *Translation, Cognition & Behavior*. 2022;5(1):1-26.
8. Rennert S. Visual Input in Simultaneous Interpreting. *Meta*. 2008;53(1):204-217.
9. Pisani E, Fantinuoli C. Measuring the Impact of Automatic Speech Recognition on Number Rendition in Simultaneous Interpreting. In: Wang C, Zheng B, eds. *Empirical Studies of Translation and Interpreting: The Post-Structuralist Approach*. 1st ed. New York: Routledge; 2021. p. 181-197.
10. Laugwitz B, Schrepp M, Held T. Construction and evaluation of a user experience questionnaire. In: Holzinger A, editor. *USAB 2008. LNCS 5298*. Berlin: Springer; 2008. p. 63-76.
11. Qualtrics. [Internet]. Provo, Utah, USA: Qualtrics; [cited 2023 June 20]. Available from: <https://www.qualtrics.com>
12. Rodríguez González E, Saeed A, Davitti E, Korybski T, Braun S. Reimagining the remote simultaneous interpreting interface to improve support for interpreters. In: Ferreiro-Vázquez Ó, Varajão A, Pereira Correia Moutinho AT, Gonçalves Araújo SL, editors. *Technological innovation put to the service of language learning, translation and interpreting: Insights from academic and professional contexts*. Berlin: Peter Lang; 2023.

13. Schrepp M, Hinderks A, Thomaschewski J. Design and evaluation of a short version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence*. 2017;4(6):103-108.
14. Saeed MA, González ER, Korybski T, Davitti E, Braun S. Connected yet distant: An experimental study into the visual needs of the interpreter in Remote Simultaneous Interpreting. In: Kurosu M, editor. *Human-Computer Interaction. User Experience and Behavior. HCII 2022. LNCS 13304*. Cham: Springer; 2022. p. 229-246.
15. Hickey S. Women take the mic in conference interpreting: Exploring the gender imbalance in conference interpreting through the lens of female practitioners. [Internet]. 2019 [cited 2023 Apr 14]. Available from: https://aiic.org/document/311/AIICBlog_Jun2019_Hickey_Women_take_the_mic_in_CI_EN.pdf

Developing a New CAI Tool for RSI Interpreters' Training: a Pilot Study

Valentina Baselli ^[1]

¹ IULM University, Milano, Italy

valentina.baselli@iulm.it

Abstract. Over the past few years, new technologies in the field of Interpreting have greatly reshaped the way interpreters work, leading to a technological turn in Simultaneous Interpreting (Fantinuoli 2018), due to the increasing use of Remote Simultaneous Interpreting (RSI) and Computer Assisted Interpreting Tools (CAI tools). When there is no human boothmate, AI-based CAI-tools are becoming “artificial boothmates” (Fantinuoli 2017), which support the interpreter before and while they deliver Simultaneous Interpreting services through automatic terminology lookup, key term identification, automatic speech recognition, real-time speech transcription, and number highlighting.

While a few researchers have investigated the field of Computer Assisted Interpreting, e. g. Fantinuoli (2017; 2018; 2019), Prandi (2018; 2020), Frittella (2022; 2023) and Defrancq (2020), more research into Computer Assisted Interpreting Training is needed, so that new technologies may be integrated into interpreter training and workflow, given their potential to help interpreters face this technological breakthrough.

This pilot study, conducted within the IULM research project “Collaboration for translation and interpreting: tools and teaching applications”, focuses on investigating the training of interpreting students on these new technologies in collaboration with the RSI-platform Converso Education by integrating the RSI-platform with a new CAI tool specifically developed for teaching purposes.

To the best of our knowledge, this RSI-platform with CAI tool specifically developed for interpreting students based on their needs is the first of its kind.

Keywords: AI-powered CAI-tool, Remote Simultaneous Interpreting (RSI), Computer Assisted Interpreter Training

1 Introduction

Over the past few years, new technologies in the field of interpreting have greatly reshaped the way interpreters work, leading to a technological turn in the sphere of Simultaneous Interpreting (Fantinuoli 2018), as the use of Remote Simultaneous Interpreting (RSI) after the Covid-19 pandemic (Baselli 2023) and Computer Assisted Interpreting tools (CAI tools) has greatly increased.

With the recent integration of CAI tools into RSI platforms, such as Kudo’s Assist and SmarTerp, the development of new tools now aims to increase the efficiency of the interpreter’s workflow and provide interpreters with a better user experience (Frittella 2023).

In 2022, we started teaching a new Remote Simultaneous Interpreting class at IULM University using the Converso Education Platform. In substance, fifty first-year and

fifty second-year students of the Master’s Degree Course in Conference Interpreting attended ten Remote Simultaneous Interpreting lessons.

Besides experiencing technical problems due to poor internet connections or inadequate equipment (such as devices and headphones), we noted that the students also ran into some difficulties during the lessons with the remote interpreting itself, especially where numerals and specialized terms were concerned. For this reason, we asked the students to complete a questionnaire on the main difficulties they encountered during Remote Simultaneous Interpreting and on useful resources to overcome those difficulties. The aim was to establish if the use of a CAI tool including an “artificial boothmate” (Fantinuoli 2017), which displays what are known as "SI problem triggers", might be helpful during RSI lessons.

2 Survey on Students’ Requirements

The goal of the survey was to explore the current students’ requirements in the field of computer assisted interpreter training, to find ways to help trainee conference interpreters face the above-mentioned technological turn and provide them with the proper tools to adequately manage RSI through the development of a new CAI tool based specifically on their requirements.

2.1 Sample

The survey was hosted on Google Forms and sent to the participants via email in April 2023. The thirty participants were regular, full-time students enrolled in the first year of the Master’s Degree Course in Conference Interpreting at IULM University. The participants’ A language was Italian, and B language was English.

2.2 Questionnaire

The user requirements for our new CAI tool stem from a questionnaire completed by thirty trainee interpreters at the end of their RSI lessons and a focus group consisting of six students conducted a few days before the recordings were made.

In the questionnaire, the students were asked to answer specific questions on their preferences related to a CAI tool developed for teaching purposes.

Table 1. Question 1: If it were possible to receive support during Simultaneous Remote Interpreting, would it be useful for you to see the numbers uttered by the speaker?

Response	Percentage of responses
Yes	97%
No	3%

Almost all the students replied that it would be useful to see the numbers uttered by the speakers on the screen and most of them (67%) preferred the numerals and punctuation to be converted into target language format.

Table 2. Question 2: If it were possible to receive support during Simultaneous Remote Interpreting, would it be useful for you to see the specialized terms uttered by the speaker?

Response	Percentage of responses
Yes	93%
No	7%

Almost all the interviewees replied that it would be useful to see the specialized terms uttered by the speakers on the screen and 90% of them would prefer to have them displayed both in the source and target languages.

Table 3. Question 3: If it were possible to receive support during Simultaneous Remote Interpreting, would it be useful for you to see the named entities (places, persons, etc.) uttered by the speaker?

Response	Percentage of responses
Yes	93%
No	7%

Among the survey respondents, almost all replied that it would be useful to see the named entities uttered by the speakers on the screen.

Table 4. Question 4: If it were possible to receive support during Simultaneous Remote Interpreting, would it be useful for you to see the entire transcription of the speech?

Response	Percentage of responses
Yes	50%
No	50%

It is interesting to note that exactly half of the students would find it useful to see the whole transcript while the other half would prefer not to see the whole running transcription of the speech uttered by the speaker, but only single elements.

2.3 Results

As most of the respondents who took part in the survey reported that displaying numbers, specialized terms, and named entities on the RSI platform during RSI classes

would be of great benefit, we decided to develop a CAI tool together with Converso, implementing it in the Converso Education Platform.

This tool is a prototype of an ASR-supported CAI tool that transcribes speech delivered in English and Italian, when enabled by the teacher on the basis of the specific needs, and automatically provides the interpreter with numerals and their unit of measurement, and translation options for terminology (drawn from a previously provided glossary). The Converso Education Platform includes a button called “view list”, which activates either automatic transcription or the CAI tool, and a “focus mode” button, which displays the transcript with highlighted specialized terms and numbers. A two-second latency was chosen, in accordance with studies conducted by Fantinuoli and Montecchio (Fantinuoli and Montecchio 2022), considering the average ear-voice span of interpreters.

Furthermore, according to the students’ preferences and Frittella’s recommendation (Frittella 2023), it was decided to let the CAI tool display numerals in their final version, without the partial rendition proposed by the ASR, as well as the numeral together with the following element in the sentence, which is usually either the referent or the unit of measurement. The above-mentioned items remain on screen for as long as there is enough room on the screen (EABM 2021). In its current state, however, our CAI tool prototype displays the numbers in the source language without showing the following element in the sentence. Adjustments will be made in the coming months before the study is conducted in the Autumn.

3 Pilot Study

3.1 Development of the CAI Tool Integrated in the RSI Platform Converso Education

According to the principles defined by Fantinuoli (Fantinuoli 2017) for ASR-based CAI tools, the final version of the CAI tool integrated in the Converso Education Platform will:

- be speaker-independent
- be able to manage continuous speech
- support large-vocabulary recognition
- support vocabulary customization for the recognition of specialized terms
- have high performance accuracy, i.e. a low word/error rate (WER)
- have high precision, i.e. fraction of relevant instances among the retrieved instances
- have high recall, i.e. the fraction of relevant instances that have been retrieved over the total amount of relevant instances present in the speech (with precision having priority over recall, in order to avoid producing results that are not useful and may distract the interpreter)
- have a distraction-free graphical user interface to present the results.

3.2 Evaluation of the CAI Tool

According to Frittella’s methodology, our ASR-supported CAI tool prototype was evaluated via tool performance, users’ performance, and users’ perception (Frittella 2023: 55).

Tool performance was assessed by adopting the same principle first used by Fantinuoli (Fantinuoli 2017), through accuracy, precision, and recall. Accuracy (i.e. word-error rate) is the percentage of wrongly displayed items (numerals, terminology, and named entities) out of all items that should have been displayed. Precision refers to the number of correct positive results divided by the number of all positive results, and recall indicates the number of correct positive results divided by the number of positive results that should have been returned.

User performance was assessed both through the rendition of individual items (interpreted specialized terms and numerals) and by considering the meaning of the interpreted part of the speech.

Users’ perception was evaluated through a post-task questionnaire.

3.3 Preliminary Test

In order to assess tool performance, a preliminary test was carried out with a pre-recorded speech and no interpreting. The aim was to evaluate the ASR precision regarding numbers and terminology. Named entities are not recognized by the CAI tool prototype at this stage.

The development of the tool has taken the principles defined by Fantinuoli (2017) for ASR-based CAI tools into consideration. Specifically, in order to be used with a CAI tool, an ASR system needs to be speaker-independent, be able to manage continuous speech, support large-vocabulary recognition and vocabulary customization for the recognition of specialized terms, and have high performance accuracy.

The table below shows the results of the preliminary test conducted on the specialized terms provided through a glossary and identified by the CAI tool. The speech (described below) was the same interpreted by the subjects during the recordings.

Table 5. Results of the preliminary test conducted on specialized terms.

Total specialized terms	Errors	Omissions
44	-	4

The four omitted specialized terms were two noun phrases consisting of five constituents (solar-thermal heating water system, natural-gas-based electricity generation), and two noun phrases containing acronyms (REs installations, Refuse-derived fuel RDF).

As far as numerals are concerned, the following table shows the results of the preliminary test conducted on the numbers identified by the CAI tool.

Table 6. Results of the preliminary test conducted on numbers.

Total numbers	Errors	Omissions
21	1	-

The numeral that was incorrectly displayed was 405, which was split into two numbers (400 and 5). The following table shows an overview of accuracy, precision and recall considering all 65 stimuli (numbers + terms). Precision and recall values are expressed from 0 to 1, with 1 being the maximum value, whereas accuracy is expressed as a percentage of error (the lower the percentage, the more accurate the result).

Table 7. Tool performance assessment

WER (accuracy)	Precision	Recall
7.7%	0.98	0.92

3.4 Pilot Study

The preliminary test and the pilot study were carried out at the Converso Hub in Milano at the beginning of June 2023. The Converso hub is Italy's first Remote Simultaneous Interpreting hub consisting of fourteen booths, a conference room and three control rooms. Every booth is fitted with professional equipment (ISO 20109:2016), ISO compliant interpreting console, a 27" full-HD display (single monitor in single-desk booths and double monitor in double-desk booths), professional gooseneck microphone or professional headset microphone, and professional headphone.

The six participants were regular, full-time students enrolled in the first year of the Master's Degree Course in Conference Interpreting at IULM University who were attending the RSI classes. The participants' A language was Italian, which is the target language of the study, and their B-language was English, the source language of the speeches.

As usability is determined according to the relevance of a product for a particular user and aim, the participants are representative of the target users (Frittella 2023: 20), that is to say Master's Degree students in Conference Interpreting.

Each test subject was asked to interpret two speeches of a similar length and lexical density on renewable energy transition. In one case, a Microsoft Word table glossary with the relevant terminology was provided. No glossary was provided in advance to help the interpreters with the second speech, but specialized terms and numbers were displayed by the CAI tool. The second speech was reinterpreted by the subjects while the entire ASR transcript of the speech was displayed. The performance of the students with ASR transcript will be compared with that obtained without CAI tool in a future study. A few days before the experiment the participants were given access to the Converso platform with transcription and CAI tool, in order to avoid the "novelty effect". However, according to the post-task questionnaire results, more in-depth training would have been useful.

The two speeches with similar difficulty level contained the same number of specialized terms and numerals. They have been prepared by the author and pre-recorded by a native American English teacher. Both are interpreter trainers. The average speed of the speeches was 100 words per minute in accordance with the indication given by Korpál and Stachowiak-Szymczak (2020) on the ideal speech rate for interpreters. The first speech (1011 words) was ten minutes and five seconds in duration, with 65 stimuli (21 numbers + 44 terms: one of which was a unigram, 19 bigrams, 20 trigrams, 2 4-grams and 2 5-grams) whereas the second speech (1000 words) was exactly ten minutes long with 65 stimuli (21 numbers + 44 terms: 9 of which were unigrams, 23 bigrams, 10 trigrams, 2 4-grams and no 5-grams, since the CAI tool did not recognize the 5-grams).

After the test, participants were asked to complete a post-task questionnaire on their perception and assessment of the tool and their preference for a display format (terms on the left and numbers on the right or vice versa, new items in a bold font, a larger font size, etc.) in addition to further comments and open questions on the use of the CAI tool.

Subsequently, the subjects' deliveries were checked for the percentage of terms and numbers translated in the first and in the second speech, which indicates whether the use of the CAI tool would help improve terminological and number precision in RSI classes. The following tables show the number of correctly translated stimuli with the support of the CAI tool and with the Microsoft Word glossary.

Table 8. Number of correctly translated stimuli with CAI tool

	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6
Numbers	61.9%	81%	95.2%	52.4%	81%	71.4%
Specialized terms	68.2%	54.5%	86.4%	75%	88.6%	63.6%
Total	66.15%	63%	89.2%	67.7%	86.15%	66.15%

When supported by the CAI tool, the six subjects correctly translated 73% of the stimuli on average. Interestingly, the numeral wrongly displayed by the CAI tool (405) was correctly interpreted by two trainees.

Table 9. Number of correctly translated stimuli without CAI tool

	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6
Numbers	38%	52.4%	66.7%	28.6%	33.3%	33.3%
Specialized terms	27.3%	43.2%	61.4%	50%	45.5%	29.5%
Total	30.8%	46.15%	63.1%	43.1%	41.5%	30.8%

Conversely, when interpreting with a Microsoft Word glossary, the six subjects correctly translated 42.6% of the stimuli on average. What clearly emerges from the data analysis is that all subjects show a remarkably better and more precise performance in RSI with CAI tool. However, some interpreted sentences in two trainees' deliveries did

not make sense. Although the numbers and specialized terms were correct when supported by the CAI tool, the segments following the numbers were sometimes mistranslated. A qualitative analysis relating to the meaning of the interpreted speech segments will be included in the future study.

Comparing the two tables, it can be observed that the number of correctly interpreted stimuli (both numerals and specialized terms) are higher in the Remote Simultaneous Interpreting with CAI tool support than in the RSI performed with Microsoft Word glossary for all six subjects (student 1: 66.15% vs 30.8%; student 2: 63% vs 46.15%; student 3: 89.2% vs 63.1%; student 4: 67.7% vs 43.1%; student 5: 86.15% vs 41.5%; and student 6: 66.15% vs 30.8%). The data from the future study on a larger sample with the adjusted CAI tool will provide a broader view of the phenomenon and produce more findings regarding the students' deliveries with or without CAI tool.

3.5 Results from the Post-Task Questionnaire

The trainees were asked to complete a post-task questionnaire after the test to evaluate user perception and satisfaction. According to the results, the 6 subjects were overall satisfied with the use of the CAI tool and emphasized that it is easy to use and intuitive, but some adjustments need to be carried out to make the CAI tool even more effective, as it is a prototype. The average scores obtained in the various categories analyzed (perceived ease of use, effectiveness, ease of learning, timeliness, dependability) range from 6.8 to 8.8, with a prevalence of an average score of 8 out of 10.

Table 10. Users' perception

Question	Average of the 6 scores (from 1 to 10)
Your satisfaction with the CAI tool's support during the test	8
The CAI tool was easy to use	8.8
The CAI tool helped me improve the accuracy of my delivery	8
No training is required to use the CAI tool effectively	6.8
The input provided by the CAI was timely	7.8
I felt that I could rely on the CAI tool's support	8

In terms of design-related recommendations, subjects have different opinions on the choice of graphic representation of the elements displayed by the CAI tool. Currently specialized terms, numbers and transcriptions are shown in a box on the right of the screen while new items appear in the same font and remain on the screen until there is no more room.

Table 11. Design-related recommendations given by the subjects.

Specialized terms	on the left 2	on the right 3	In the center 1
Numbers	on the left 1	on the right 3	In the center 2
New items	bold font 5	larger font size 0	No font change 1

Moreover, it is interesting to note that 5 out of 6 stated that they perceived the CAI tool as being more reliable than a human boothmate in terms of precision and speed in suggesting specialized terms and numbers. On the one hand, it emerged that for a trainee it is reassuring to know that the interpreter has not to search for terminology through a glossary, even if a human boothmate psychologically helps make the interpreter feel less alone and can better understand the interpreters' needs. On the other hand, a trainee found the transcript and particularly the stimuli moving on the screen very distracting. It would be interesting to see if the same difficulties would emerge after more training sessions with the CAI tool.

4 Conclusions and Future Work

This test is a pilot study for a broader research project of the “International Center for Research on Collaborative Translation - IULM” that involves the development of a CAI tool integrated into the Converso Education RSI Platform and the usability test. The findings of this pilot study will help us redesign the broader study which is scheduled for Autumn 2023 on a larger sample (between thirty and fifty students enrolled in the second year of the Master’s Degree Course in Conference Interpreting at IULM University). Furthermore, the future study will investigate ASR output in the form of both “short prompts” and full ASR transcript, since the author has also decided to record the students’ performance with the entire speech transcript. This comparison might lead to more findings which may complement the existing insights from Fantinuoli, Frittella and Prandi. The results of this pilot study are not intended to be exhaustive but form the basis for in-depth research work on Computer Assisted Interpreting Training.

Acknowledgment

The author wishes to thank Converso for developing the CAI tool.

References

1. Baselli, V.: La traduzione collaborativa in interpretazione: il connubio tra interprete e tecnologia. In: Laurenti, F. (ed.), *La traduzione collaborativa tra didattica e mercato globale delle lingue*, Aracne, Roma (2022).
2. Baselli, V.: L'Interpretazione Simultanea da Remoto: stato attuale, sfide e prospettive. In: Fuehrlinger, S., Proietti, P. (eds.), *Transiti linguistici - Traduzioni, interpretazioni, mediazioni linguistiche*, Aracne, Roma (2023).
3. Converso Homepage, <https://converso.cloud/en/> last accessed 2023/4/20.
4. Defrancq, B., Fantinuoli, C.: Automatic Speech Recognition in the booth: Assessment of system performance, interpreters' performances and interactions in the context of numbers. *Target*, DOI (2020).
5. Ergonomics for the Artificial Booth Mate (EABM) <https://www.eabm.ugent.be/survey/> (2021).
6. Fantinuoli, C.: Speech recognition in the interpreter workstation. In: Esteves-Ferreira, J., Macan, J., Mitkov, R., Stefanov, O. (eds.), *Proceedings of the Translating and the Computer 39*, 25–34. *AsLing*, London (2017).
7. Fantinuoli, C.: Interpreting and technology: The upcoming technological turn. *Interpreting and technology*, Language Science Press, Berlin (2018a).
8. Fantinuoli, C.: Computer-assisted interpreting: challenges and future perspectives. In: Durán, I., Corpas, G. (eds.), *Trends in e-tools and resources for translators and interpreters*. 153-174, Leiden, Brill (2018b).
9. Fantinuoli, C.: The technological turn in interpreting: the challenges that lie ahead. In: *Proceedings of the conference Übersetzen und Dolmetschen 4.0. - Neue Wege im digitalen Zeitalter*, 334-354 (2019).
10. Fantinuoli, C., Prandi, B.: Teaching information and communication technologies: a proposal for the interpreting classroom. *Journal of Translation and Technical Communication Research*, 11(2), 162-182. (2018).
11. Fantinuoli, C., Montecchio, M.: Defining maximum acceptable latency of AI-enhanced CAI tools. In: *Proceedings of LingTech21* (2022).
12. Frittella, F.: CAI Tool-Supported SI of Numbers: A Theoretical and Methodological Contribution, *International Journal of Interpreter Education: Vol. 14: Iss. 1, Article 5.* (2022).
13. Frittella, F.: Usability research for interpreter-centred technology: The case study of SmarT-erp. (*Translation and Multilingual Natural Language Processing 21*). Language Science Press, Berlin (2023).
14. Korpál, P., Stachowiak-Szymczak, K.: Combined problem triggers in simultaneous interpreting: exploring the effect of delivery rate on processing and rendering numbers. *Perspectives Perspectives: Studies in Translation Theory and Practice*, 28 (1): 126–143, <https://doi.org/10.1080/0907676X.2019.1628285> (2020).
15. Prandi, B.: Designing a Multimethod Study on the Use of CAI Tools during Simultaneous Interpreting. In: *Proceedings of the 39th Conference Translating and the Computer*. *Translating and the Computer*, London (2017).
16. Prandi, B.: An exploratory study on CAI tools in Simultaneous Interpreting: theoretical framework and stimulus validation. In: Fantinuoli, C. (ed.) *Interpreting and technology*, pp. 28–59, Language Science Press, Berlin (2018).
17. Prandi, B.: CAI tools in interpreter training: where are we now and where do we go from here?, in *TRAlinea* (2020).

Introducing Speech Recognition in Non-live Subtitling to Enhance the Subtiter Experience

Željko Radić [0009-0003-9451-5557],
Sabine Braun [0000-0002-6187-3812] and Elena Davitti [0000-0002-7156-9275]
University of Surrey, Guildford, Surrey GU2 7XH, UK
z.radic@surrey.ac.uk

Abstract. Interlingual Subtitle Voicing (ISV) is a new technique that focuses on using speech recognition (SR), rather than traditional keyboard-based techniques for the creation of non-live subtitles. SR has successfully been incorporated into intralingual live subtitling environments for the purposes of accessibility in major languages (real-time subtitles for the deaf and hard of hearing). However, it has not yet been integrated as a helpful tool for the translation of non-live subtitles to any great and meaningful extent, especially for lower resourced languages like Croatian. This paper presents selected results from a larger PhD study entitled ‘Interlingual Subtitle Voicing: A New Technique for the Creation of Interlingual Subtitles, A Case Study in Croatian’. More specifically, the paper focuses on the second supporting research question that explores participants feedback about the ISV technique, as a novel workflow element, and the accompanying technology. To explore this technique, purpose-made subtitling software was created, namely SpeakSubz. The constant enhancements of the tool akin to software updates are informed by participants’ empirical results and qualitative feedback and shaped by subtitlers’ needs. Some of the feedback from the main ISV study is presented in this paper.

Keywords: ISV, Interlingual Subtitles, Non-live Subtitles, Subtiter Experience, SUBX, SpeakSubz, Human-Machine Interaction, Action Research.

1 Introduction

All too often subtitling tools and, consequently, workflows are determined by clients, be they translation agencies as an intermediate link in the workflow or end clients. Subtitlers themselves always need to adapt to the existing workflows and use tools mandated by clients. Nowadays these are mostly online proprietary subtitling tools all of which differ from one another and offer various functionalities. As a result, subtitlers are expected to learn how to use multiple tools to remain competitive in the market. Unfortunately, when developing such software some of the most important agents in the subtitling workflow, subtitlers themselves, are rarely being consulted. The Interlingual Subtitle Voicing (henceforth ISV) study aims to amend this by adopting a subtitler-centric approach to the investigation of a new subtitling workflow.

The topic presented in this paper is a part of a larger doctoral study which explores the integration of speech recognition (SR) into a non-live subtitling workflow using a newly introduced technique. The ISV study generated empirical data on efficiency, measured as a combination of translation quality (from English into Croatian) and SR accuracy in Croatian, as well as the duration of various workflow elements, such as typing, voicing and editing. However, this paper will focus on the qualitative part of the study that strives to capture the subtitler experience (SUBX) with this new form of human-machine interaction through the feedback provided by participants. This feedback will inform subsequent updates of the ISV technique and the specialised software (SpeakSubz), created specifically to implement it. The ultimate goal is creating a customisable tool with a plethora of functionalities that can enhance subtitlers' workflows and work environments based on their needs and preferences.

2 The Interlingual Subtitle Voicing Technique

ISV is a newly devised hybrid technique drawing from three established modes of language transfer: subtitling, interpreting and respeaking. The technique is meant for the translation of pre-recorded or non-live subtitles and leads to a workflow that differs from traditional subtitling workflows. It does so by introducing a new variable into the process: voicing, known as respeaking in live subtitling. This type of human-machine interaction includes a subtitler, or a voicer, translating subtitles from a source language (SL) by speaking into a microphone connected to SR software. This specialised software turns vocal input in a target language (TL) into textual output in TL in the form of subtitles. The process is similar to interlingual respeaking, without the element of immediacy since the ISV technique occurs in non-live environments.

Subtitling is a crucial component of the ISV technique. The traditional subtitling workflow typically involves a fairly linear sequence of activities. These are often determined by the client or software used to produce subtitles, especially in cloud-based environments. Within that workflow, subtitlers have some leeway in how they handle different workflow components. For the purposes of this paper, it is relevant to distinguish these two levels of workflows, which can be referred to as macro-level and micro-level workflows, respectively. The term is drawing on the distinction made by Alves et al. [1] who distinguish between macro and micro translation units.

In this paper we consider workflows coming from an end client or an intermediary translation agency as a macro-workflow that subtitlers normally cannot influence, such as the use of proprietary subtitling tools, linguistic and technical guidelines, deadlines etc. On the other hand, each subtitler has his or her own individual translation strategies, which we label as micro-workflows: these mostly have to do with the order in which a subtitler performs certain tasks and strategies s/he uses within the established macro-workflow.

Figure 1 below shows the basic workflow configuration devised for the purposes of the ISV study that can, as needed, be adapted to individual subtitlers' micro-workflows in the future. This workflow was also devised with the ISV study in mind in order to facilitate empirical measurements related to duration of various stages of a subtitling process and make participants aware of these various workflow elements.

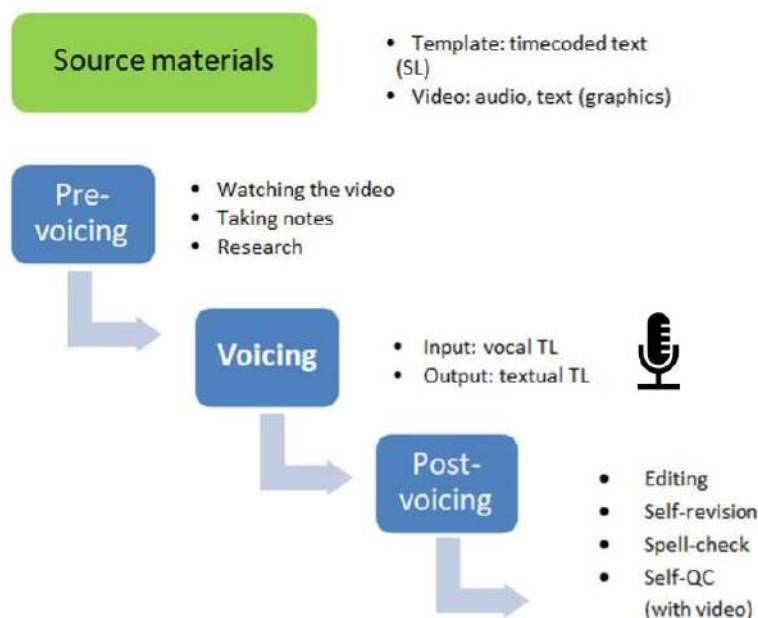


Fig. 1. The ISV workflow consists of three stages: a pre-voicing, a voicing and a post-voicing stage.

The pre-voicing stage of the ISV workflow includes preparatory activities that occur before the translation/voicing process. The central voicing stage includes reading the template (subtitles in a source language), mentally translating the source text into a target language and speaking it out into SR software, including punctuation, while monitoring the textual output in a target language and reacting as needed. The post-voicing stage includes various edits and checks that need to be conducted to ensure the desired maximum accuracy of 100%, i.e., broadcast-ready subtitles. However, these are a part of a macro-workflow and as such not relevant for the ISV technique.

3 Related Work

The ISV technique was developed with subtitlers in mind. Such subtitler-centric research in academia is extremely rare and current studies have focused either on the final product, (i.e., subtitles), the way they are created, (i.e., the process) or the target audience. The existing scholarship ranges from research about traditional subtitling [2–4], quality of subtitles through practical experiments [5], as well as quality of templates [6] or theoretical quality assessment models [7]. More recent research has delved into media accessibility [8–10], new subtitling workflows, especially those that involve machine translation [8] and post-editing [9] and the translation process itself, be it traditional translation process research or TPR [10] or on online platforms [11]. Finally,

there are also audience reception studies [12, 13] but almost no studies that focus on subtitlers' various micro-workflows and their evaluation of traditional or new hybrid workflows and, most importantly, their agency in creating the tools they might use in the future. The ISV technique aims to address this gap and contribute to the existing literature.

As already mentioned, the ISV research focuses on a new hybrid workflow implemented via a custom made proprietary online platform with an emphasis on the impact of integrating SR as a new major variable in the subtitler experience. To explore this workflow, subtitlers' workspaces and needs, the current study relies on two main frameworks: action research (AR) and translator experience (TX). On one hand, AR [14–16] allows for the study of a new hybrid workflow and observes how it affects efficiency in terms of both speed and accuracy. On the other hand, the notion of TX [17], as well as studies about multimodality in translator-computer interaction [18–20] have informed the development and testing of ISV. This paved the way for the notion of Subtitler Experience (SUBX), which not only examines subtitler themselves, but also their work processes, work environments and their evaluation of suggested techniques and workflows.

4 Methodology

This paper focuses on the second supporting research question of the ISV study: how does the ISV technique influence SUBX in a traditional subtitling environment, primarily in terms of ergonomics? We present selected findings from the qualitative portion of the study that explored the SUBX. This paper presents how participants responded and adapted to this new hybrid workflow and how they evaluated this novel form of human-machine interaction.

15 participants, Croatian native speakers, predominantly experienced subtitlers (13), took part in the ISV experiment, which included a 9-hour online guided self-training to be completed within a one-month window. All the necessary materials (SL subtitles, videos, the technique presentations, video tutorials) and tools needed for the experiment (integrated in the specialised tool SpeakSubz) were made available to participants on a website created for this study. The experiment had to be carried out entirely online because of the Covid-19 pandemic. During the ISV training, participants gradually learned voicing as a new skill and they progressively developed their abilities through a series of exercises, leading up to the final two exercises (7 & 8) that were quantitatively analysed.

To evaluate efficiency of the ISV technique and its underlying technologies, which was the focus of the first supporting research question, two variables were deemed relevant. Firstly, the speed of traditional typing compared to voicing, with and without editing. Secondly, Croatian SR accuracy was measured to gauge how close it was to the industry standard of 98% for live intralingual subtitling, while keeping in mind that ISV is meant for non-live subtitling. Moreover, there is ample time to correct SR errors in the final version of the translation. In addition, there is currently no established accuracy benchmark for interlingual respeaking, and this metric has yet to be explored

and consolidated. This quantitative data related to accuracy and speed will be published in future papers upon the completion of the PhD thesis,

The most relevant datasets for this paper were the post-testing questionnaires and optional follow-up interviews. The questionnaires consisting of 30 questions were delivered through a Google form embedded on the ISV website¹ and were completed by 15 participants. The questionnaire was divided into 5 sections: ISV Presentation, Croatian Speech Recognition and SpeakSubz, Translation Workflow and Skills, Translation Hardware and Future Translation Work.

Moreover, participants had the opportunity to apply for optional follow-up interviews (online, under 45 minutes) to provide a more in-depth evaluation of the technique upon completing their ISV training and testing. Six participants chose to do so, and despite the limited number, a wealth of information was gathered about five different topics: the ISV website, the ISV technique, specialized software SpeakSubz, participants' workflows and workspaces and, finally, future translation practices.

Additionally, voice and screen recordings of participants' performance in the two final exercises, initially set up for tracking durations of various processes, were used as a secondary source of data. This data gave a glimpse into participants' behaviour during the voicing, typing and editing processes. These evaluations and general feedback, as well as voice and screen recordings, enabled exploration of human-machine interaction in non-live subtitling and how it can be further improved. Due to space limitations of this paper, the following section will only present selected findings about subtitlers' experience with the ISV technique, while forthcoming subsequent qualitative and quantitative data will give a more complete picture of ISV and especially the SUBX.

5 Selected Findings

The selected results presented in this section come from the post-testing questionnaires and optional interviews conducted over Skype, transcribed and thematically analysed. For the participants in the ISV experiment, that was the first time they had encountered voicing professionally. Although the empirical data of the ISV study showed that Croatian SR is not yet on par with major languages such as English, Spanish, German or Italian, the results from the post-testing questionnaires and optional interviews indicate that participants were intrigued and often pleasantly surprised by the new technique.

For example, in Question 8 of the post-testing questionnaire, participants' subjective evaluation of SR accuracy showed that they believe that the underlying technology is not yet perfect, which was also to some extent confirmed by the empirical data of the study. Most participants rated Croatian SR accuracy with a 3, with the median mark being 3.2 out of 5 (Figure 2).

¹ <https://isvresearch.eu/>

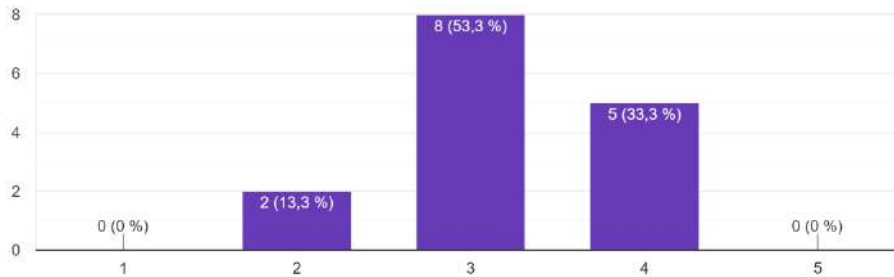


Fig. 2. Question 8: Participants' subjective evaluation of the accuracy of Croatian speech recognition.

Nevertheless, the rating was more positive when it came to their experience of working with Croatian SR (Question 7). The majority of participants rated it positively, with 10 participants rating it either 4 or 5, while only one participant rated it with a 2 (Figure 3). The median value was 4.06 and much more favourable than in Question 8. This suggests that participants mostly enjoyed working with Croatian SR, despite the imperfections of the underlying technology and relatively short training.

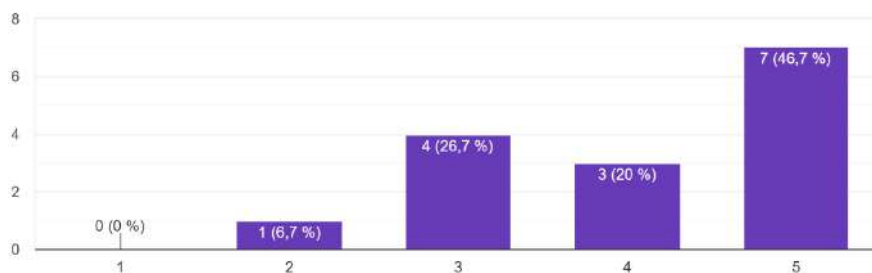


Fig. 3. Question 7: Participants' experience of using Croatian speech recognition.

The participants' feedback regarding the specialised software developed for this study (Question 9) was even more positive. As shown in Figure 4, the mean value amounted to 4.13 out of 5. This finding was also confirmed in Question 11 of the questionnaire and optional follow-up interviews, where participants described the software to be "user-friendly", "easy to use", "visually appealing" and even "game-like". They also enjoyed its various functionalities such as reading speed markers, the comment section, pre-loaded tasks and voicing itself. Participants also suggested what other functionalities could be added in the future (Question 12) such as: automatic measurement units converter, an undo button and machine translation (MT). The latter was in the meantime added to the software and is currently in the beta phase ready for wider testing.

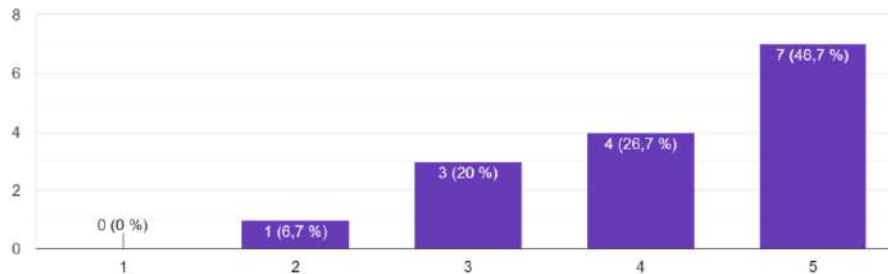


Fig. 4. Question 9: Participants' experience of using SpeakSubz.

When asked whether participants had difficulties and what they were (Questions 5 and 6), some participants found certain exercises had too many steps in the workflow and because of that they needed to consult reminders. These reminders were in the form of “Read Me” textual files within each exercise. Other participants, on the other hand, were not always sure what to record and when. However, some participants in the follow-up interviews stated that they got used to the workflow by the final two exercises and the process became easier for them as the training progressed. Some participants also experienced some slight technical difficulties, with their voice not being captured or the recording functionality not working properly. Technical issues have been either resolved and/or put under the FAQ section of the website to clarify how to solve these issues in the future.

Regarding the comparison between voicing and typing, some participants did find voicing to be faster whereas others believed they were still faster when typing but nevertheless see the potential of the ISV technique. One participant, a highly experienced subtitler, also suggested that the technique might be better suited for interpreters since subtitlers are not used to voicing their translations and might need longer to become accustomed to it.

However, despite the potential shortcomings of the technique described above, in Question 17, 60% of participants stated they would like to continue developing their voicing skills in the future (Figure 5) and 33.3% indicated they might consider it. Taking into account the imperfections of the SR technology, the complexity of the technique, the short duration of the guided self-training and some minor bugs in the software that need to be corrected in the next version, this is a very positive result. Besides the need for training “voicers” in the future, it is encouraging to see that most participants enjoyed working with the technique, had a positive SUBX and would like to continue to use the technique in the future.

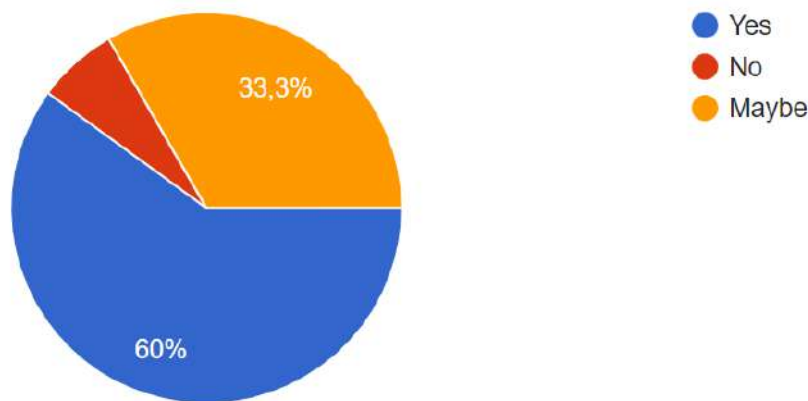


Fig. 5. Question 17: Future training in voicing.

6 Future Research

After establishing the accuracy of Croatian SR within overall quality evaluation and durations of various subtitling processes to determine efficiency, it is promising to note positive results in terms of participants' experience and reception. The future success of the ISV technique and specialised software as well as their usability in education and the audiovisual translation (AVT) industry will depend on software upgrades and technique adaptations. Experience from the AVT industry shows that machine translation seems to be increasingly used in subtitling workflows. Additionally, based on suggestions from some of the participants already working with it on a daily basis, adding MT functionality within the ISV technique would be the next logical step. This hybrid approach could populate subtitle boxes with MT (subtitle by subtitle or all subtitles at once) and subtitlers could use voicing to correct errors, thus reducing the need for typing with physical keyboards. That workflow configuration will be tested next at Croatian universities as of 2024/2025.

The technique and the software itself could be further developed in two different ways: one for academia and one for the AVT industry. SpeakSubz is not only a practical tool for professional subtitlers, but also a training and a methodological tool for students. As already mentioned, Croatian universities have already shown interest in using this tool and the technique in the training of subtitlers. The tool can be used, for example, to teach AVT students the basics of subtitling, text reduction (reading speed markers), line breaks ("new line" virtual button), voicing (respeaking) and MT post-editing.

When it comes to SUBX, a key element that needs to be studied in the next cycle of ISV research is the usability of the technique and software in real-life professional situations, especially on handheld and touchscreen devices to widen and enhance the SUBX in terms of ergonomics. While most AVT-related research nowadays is conducted in controlled media lab conditions, it is important to capture the use of the technique in subtitlers' real workspaces which nowadays, more often than not, are their own homes. The ISV methodology allows for anonymous research in real-life professional

situations, without intruding on participants' privacy. This methodology could be developed further to include a more meaningful utilisation of mobile and handheld devices in the next phase of the study, with a possible longitudinal study to track participants' progress.

7 Conclusion

The ultimate goal of this study is to develop training in a hybrid translation mode (ISV) and offer users, intended as language professionals at large, customisable tools with multiple functionalities to optimise their everyday work. Unlike existing subtitling tools, users can directly influence the functionality of SpeakSubz by suggesting features that would help them work more efficiently and can be implemented in a timely manner. In addition, subtitlers can also evaluate these novel and hybrid workflows, thus giving them agency that is rarely present in subtitling software development, whereby future researchers could rapidly replicate or adapt them for their own research purposes.

As underlying technologies advance, ISV could potentially be used for live subtitling into Croatian as well as to enhance accessibility. The technique could also be applied and/or adapted to other lower-resourced languages which are lagging behind the major languages simply because the technology is not yet available to them. Rather than waiting for technology to be fully ready, proactive measures can be taken by researchers to improve existing ones through human input, as is the case in ISV. ISV can hopefully lead the way in this type of subtitler-centric research and ensure professionals are trained to live up to the challenges of new SR and MT reality both in AVT studies as well as in the AVT industry. Participants' positive experience with the ISV technique and positive evaluation of SUBX certainly encourages us to continue with this strand of research in the future.

References

1. Alves, F., Pagano, A., Neumann, S.: Translation units and grammatical shifts. Towards an integration of product-and process-based translation research. *Translation and cognition*. (2010).
2. Ivarsson, J.: Code of Good Subtitling Practice. *Subtitling and Translation*. (2010).
3. Ivarsson, J., Carroll, M.: Code of Good Subtitling Practice. *Subtitling*. (1998).
4. Cintas, J.D., Remael, A.: *Audiovisual Translation: Subtitling*. (2014). <https://doi.org/10.4324/9781315759678>.
5. Szarkowska, A., Díaz Cintas, J., Gerber-Morón, O.: Quality is in the eye of the stakeholders: what do professional subtitlers and viewers think about subtitling? *Univers Access Inf Soc*. 20, (2021). <https://doi.org/10.1007/s10209-020-00739-2>.
6. Oziemblewska, M., Szarkowska, A.: The quality of templates in subtitling. A survey on current market practices and changing subtitler competences.

- Perspectives: Studies in Translation Theory and Practice. 30, (2022). <https://doi.org/10.1080/0907676X.2020.1791919>.
7. Pedersen, J.: The FAR model: Assessing quality in interlingual subtitling. *Journal of Specialised Translation*. (2017).
 8. Bywood, L., Georgakopoulou, P., Etchegoyhen, T.: Embracing the threat: machine translation as a solution for subtitling. *Perspectives: Studies in Translatology*. 25, (2017). <https://doi.org/10.1080/0907676X.2017.1291695>.
 9. Koponen, M., Sulubacak, U., Vitikainen, K., Tiedemann, J.: {MT} for subtitling: User evaluation of post-editing productivity. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. (2020).
 10. Orrego-Carmona, D.: Using translation process research to explore the creation of subtitles: an eye-tracking study ... *Journal of Specialised Translation*. (2018).
 11. Bolaños García-Escribano, A., Díaz-Cintas, J., Massidda, S.: Subtitlers on the Cloud: The Use of Professional Web-based Systems in Subtitling Practice and Training. *Tradumàtica: tecnologies de la traducció*. 1–21 (2021). <https://doi.org/10.5565/rev/tradumatica.276>.
 12. Božović, P.: How should culture be rendered in subtitling and dubbing? A reception study on preferences and attitudes of end-users. *Babel. Revue internationale de la traduction / International Journal of Translation*. 65, (2019).
 13. Kuscu-Ozbudak, S.: The role of subtitling on Netflix: an audience study. *Perspectives: Studies in Translation Theory and Practice*. 30, (2022). <https://doi.org/10.1080/0907676X.2020.1854794>.
 14. Cravo, A., Neves, J.: Action research in translation studies. *Journal of Specialised Translation*. (2007).
 15. Kredens, K., Bogucki, L.: Josélia Neves. In: *Perspectives on Audiovisual Translation* (2016). <https://doi.org/10.3726/978-3-653-00400-7/11>.
 16. Massey, G.: Applied translation studies and transdisciplinary action research: Understanding, learning and transforming translation in professional contexts. *Russian Journal of Linguistics*. 25, (2021). <https://doi.org/10.22363/2687-0088-2021-25-2-443-461>.
 17. Zapata, J.: Translators in the Loop: Observing and Analyzing the Translator Experience with Multimodal Interfaces for Interactive Translation Dictation Environment Design, (2016).
 18. Zapata, J.: Exploring multimodality for translator-computer interaction. In: *ICMI 2014 - Proceedings of the 2014 International Conference on Multimodal Interaction* (2014). <https://doi.org/10.1145/2663204.2666280>.
 19. Zapata, J.: Translating On the Go? Investigating the Potential of Multimodal Mobile Devices for Interactive Translation Dictation. *Tradumàtica: tecnologies de la traducció*. (2016). <https://doi.org/10.5565/rev/tradumatica.180>.
 20. Zapata, J., Kirkedal, A.S.: Assessing the performance of automatic speech recognition systems when used by native and non-native speakers of three major languages in dictation workflows. In: *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015* (2015).

Towards a Decentralize Solution for Copyrights Management in Audiovisual Translation and Media Accessibility

Serrat-Roozen Iris¹[0000-0002-7993-9234] and Oncins Estella²[0000-0002-0291-3036]

¹ Universidad Internacional de Valencia, Valencia 46002, Spain

² Universitat Autònoma de Barcelona, Bellaterra, Spain
iserrat@universidadviu.com
estella.oncins@uab.cat

Abstract. With the development of new technologies in the audiovisual sector, significant changes are taking place in the way information is processed, distributed and accessed. In this regard, blockchain technology is undoubtedly at the epicentre of the technological revolution and, despite its undeniable application in different industries, it seems to remain ignored in some academic fields, particularly in Translation Studies. This technology can be used for various purposes in our field —translating data in blocks, creating a more transparent and secure workflow in the translation process, tracking translation quality— as well as to address copyright issues and to rethink the ways in which we use, reuse, distribute and monetise the content we create.

This paper addresses two key issues in the digital media industry, namely blockchain technology and intellectual property rights management, and presents an intellectual property rights (IPR) management tool developed as part of the MediaVerse project. In addition, we will analyse the results of two focus groups conducted to validate the effectiveness of this tool among audiovisual translators and media accessibility professionals. By exploring these critical issues and demonstrating the benefits of the IPR management tool, we aim to contribute to the ongoing discourse on digital media accessibility and its importance in the current media landscape.

Keywords: Blockchain Technology, Copyright Management, Audiovisual Translation, Media Accessibility

1 Introduction

The media industry has traditionally been characterized by a high degree of centralization. A rather small number of large companies and media platforms have significant control over the market and impose their conditions on consumers, prosumers and content creators [2]. The European Union's Digital Markets Act¹ therefore identifies them as 'gatekeepers' and seeks to ensure that they operate fairly online. The implications of this centralized structure are broad and have shaped the

¹ <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R1925>

media landscape in significant ways. For instance, it has led to the concentration of ownership and editorial control, as well as the standardization of content and formats, which can limit diversity and innovation in the industry. Moreover, it has also raised concerns about the potential for media companies to have a strong influence over public opinion and democratic processes. Hence, there is a need to promote diversity, pluralism, and democratic values within the media industry. This can lead to a more open, transparent, and inclusive media environment. One that reflects the diverse voices and perspectives of society, while meeting the needs and interests of all participants in the digital ecosystem.

The internet has evolved from the read-only Web 1.0 to the interactive Web 2.0, where users generate content and participate in online communities. However, big social media platforms such as Google, Facebook or YouTube dominate the web and control issues such as data privacy, censorship, and user-generated content. The next step is Web 3.0, also known as the decentralized web [10], which emphasizes the importance of ownership in the digital space. In Web 3.0, users have full control over their data and can monetize their online activity without the need for intermediaries.

The digital world is being transformed by the rise of new digital technologies, such as blockchain and Web 3.0 technologies [10]. A world where digital assets (text, video, photos, social media accounts, software programs) are important and valuable. Consequently, digital assets need to be properly stored and well secured. This shift towards decentralization presents new challenges for managing ownership of digital assets, including IPR [1]. Furthermore, the open and collaborative nature of Web 3.0 means that digital assets can be easily shared, used, and reused, which leads to the creation of derivative works that have to be properly managed. There is a need for new models of ownership and licensing enabling content creators to manage their IPR. The MediaVerse project aims at proofing the viability of a platform that offers creators the possibility to manage their digital assets and create or co-create content.

The following article will first describe the main issues related to Blockchain technology and digital assets. Second, IPR within audiovisual translation (AVT) and accessibility services, such as Subtitles for the Deaf and Hard of Hearing (SDH) and audio description (AD) will be addressed. Third the Blockchain technology developed in the MediaVerse platform to manage copyrights will be outlined. And finally, the results gathered from two focus groups with professionals in the AVT and media accessibility field will be reported. The aim of the focus group was to evaluate to which extent the copyright management envisaged under the MediaVerse project addresses the needs of professional in the AVT and media accessibility (MA) field.

2 Audiovisual Industry and Blockchain Technology

The audiovisual industry is facing several issues that entail proper attention as we move towards the future. One of the most significant challenges is the single point of failure that arises due to centralization in the digital industry. It occurs when one part of the system fails, leading to failure of the entire system. Another major concern is the low profits earned by creators, largely due to monopolistic pricing models and the number of intermediaries involved in the process.

Likewise, the management of copyrights and licensing contracts is complex and opaque, leading to a lack of transparency. This makes it challenging for right owners to effectively manage their works online, which often leads to piracy and infringement. The limitations of Digital Right Management (DRM) technologies should also be tackled, as access from other countries is not equally established or can be restricted by licensing, and new forms of payments such as micropayments and pay-per-use content are often not considered.

Furthermore, there is an absence of a global and verified register for intellectual property (IP), and records are stored in diverse systems across the media value chain. To address these issues, it is crucial to develop innovative solutions that increase transparency, reduce intermediaries, and ensure fair return for creators along with a global register to enable easy access to ownership information.

Blockchain technology provides a way to distribute content in a secure and decentralised way, without the need for intermediaries, third-parties. It is a tamper-proof system to store and share data, allowing content creators to distribute their work in their own terms, obtaining recognition (authorship) of the assets they create, regaining control (ownership) from central platforms and keeping track of changes within a chain.

One of the most significant benefits of blockchain technology is its ability to create immutable records of transactions in a transparent way [3]. This allows every node (computer) to trace their content. In the case of copyright works, it means that ownership and usage rights can be securely registered and traced on a blockchain network.

Integrating blockchain technology in IPR management can make it more efficient, cost-effective, and fair for everyone involved. In traditional centralized systems, ownership of digital assets is often managed by centralized entities. By using blockchain technology, content creators and consumers have more direct relationships, and profits can be distributed fairly. Blockchain is one of the technologies behind MediaVerse and its IPR management system.

3 Intellectual Property Rights and Translations

At an international level, there is a broad copyright legal framework in place under which, translations are generally referred to as derivative works. Derivative works are literary and artistic works which are based on pre-existing works that are altered. In addition, Article 2(3) of the Berne Convention provides the following explanation: “Translations, adaptations, arrangements of music and other alterations of a literary or artistic work shall be protected as original works without prejudice to the copyright in the original work”. Therefore, it seems essential to determine how such pre-existing and/or original works are protected under copyright law(s), as such protection could limit their use, and the right to translate and/or to store them in a database.

This copyright legislative framework is aimed at protecting the rights of all authors in the creative industries, to which translators also belong. According to Tong King [8], there is the recursive scaling of authorship in copyright (see Figure 1). As it can be observed, in order to create a translation, authorization must be obtained from the original owner of the rights (tier 1), and in order to exploit any translation (tier 2), the

same applies. Each tier generates new copyrights for new authors, which are interlinked with the pre-existing work and should be traced. This traceability avoids the lack of transparency in tracking who creates the content or derivative works, and helps to raise visibility and awareness of one's own and others' creations.

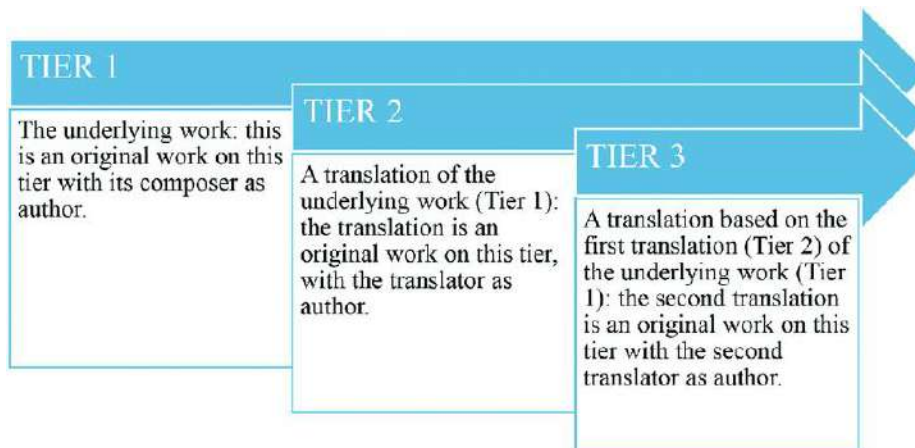


Fig. 1. The recursive scaling of originality/authorship in copyright. Source: Tong King (2020:253).

However, right management can be difficult to trace when a work is created by more than one person [4-6]. In addition, some types of translation, such as the different modalities in the AVT field (i.e. subtitles), might remain in an uncertain field, as there is no international copyright law that protects subtitles worldwide [6]. However, the digital assets that AVT and media accessibility professionals produce have the right to be properly protected.

4 Copyrights Management in AVT and Media Accessibility

Even if there is an international legal framework on copyrights management, each EU country has its own copyright laws, and an agency to manage the rights. For instance, in Spain, the Intellectual Property has been protected by a specific law since 1847 (Literary Property Act) [7]. The current Intellectual Property Law (LPI) by its acronym in Spanish) 1/1996 (BOE-A-1996-8930) has been substantially modified by various laws in order to incorporate the different European Directives in this subject. The last reform was in 2021 in order to transpose the European Directives 2019/789 and 2019/790.

In the specific case of the audiovisual translation modalities, the agency DAMA (Derechos de Autor de Medios Audiovisuales) is in charge of managing the corresponding copyrights rights. This agency is a member of the worldwide agency CISAC (International Confederation of Societies of Authors and Composers) who represent 228 member societies in 119 countries worldwide.

Author's rights that emanate from Intellectual Property are known as 'copyright'. There are two different types of copyright: moral rights and economic rights. Moral rights are personal, indisputable, and inalienable. Some of these rights exist in perpetuity: for example, the right of the author to be acknowledged and the right to the integrity of works continue forever. Economic rights and rights of exploitation are transferable rights.

Recognition of the authorship and integrity of works are the most important moral rights, together with the right to share this work. Economic rights and rights of exploitation do have time limits. Their duration depends on the legislation in each country (i.e. in Spain, they last for the lifespan of the author and following the death of the same they belong to the inheritors for a further 70 years). On the other side, economic rights and rights of exploitation are transferable rights, which means that they can be sold, ceded or shared with third parties, whether this is for economic purposes or not. Hence, the ownership of rights of exploitation does not always belong to the author, since the latter may have ceded or sold these to a third party or organization, such as an editor or compiler.

Authors may also set their own limits, enabling them to have a greater level of control over the rights of their works and to manage these more easily. This gives both, users and creators, greater access to works and enables them to be used or shared with fewer restrictions. Among the free licenses, there are also some licenses known as 'Copyleft' which are widely used in computer science. These guarantee the right of any individual to use, modify and redistribute a work as long as they share the derivative works that they create under the same, or a similar, license.

The most common free licences are the Creative Commons (CC) licenses. A set of standardized, legal tools that enable creators to share their work with others while retaining certain rights, such as the right to attribution, the right to control how their work is used, or the right to determine whether others can create derivative works from their original work. CC licenses are designed to be flexible and can be customized to suit authors' needs helping to promote legal and ethical sharing and reuse of digital assets in an open ecosystem. They can be applied to any type of creative content and have been integrated as part of the MediaVerse IPR management tool.

5 Copyrights Management in the MediaVerse Platform

In order to address issues related to intellectual property and copyright management in the audiovisual sector, the project MediaVerse², which is an H2020 Innovation Project co-financed by the EU, has developed a multimedia co-creation platform. It is an open decentralized platform with strong Intellectual Property (IP) protection that enables creators and media assets owners to create, upload and share their media while managing their intellectual property rights. MediaVerse is a proof of concept project that aims to test the viability of the platform. It gives power back to content creators (artist, audiovisual translators, freelance creators, citizen journalists, all sorts of content creators) being able to take part in the decision-making process when distributing assets and getting recognition (Creative Commons licences) for the content they create. It

² <https://mediaverse-project.eu/>

enables creators to decide how to share, distribute and monetise their content using smart contracts.

The platform includes several innovative technologies and provides a set of tools and services allowing users and creators to navigate in this new digital media industry. It offers co-creation tools where multiple users can work on their projects together including support for immersive media, such as interactive 360° videos and 3D objects; social analytics tools; a decentralised network to share the media; AI supported tools for content analysis to facilitate the process of finding suitable content fragments to build your media, and also to identify inappropriate content to protect your audience, and Automated language translation along other tools to facilitate the creation of accessible media.

The MediaVerse project has identified seven blockchain based solutions that could help content creators to address the main challenges related to copyright management in the audiovisual sector.

- Decentralized digital content ecosystem: power and ownership return to creators.
- New pricing options: new options for creators to earn by selling content.
- Monetization of content: content creators can establish direct relationships with customers.
- Distribution of royalty payments: near real time payments based on smart contracts.
- From DRM (Digital Rights Management) to smart contract: Transparent and "self - execute" right management underlying smart contracts.
- Attribution: Blockchain increases the visibility and availability of the information regarding copyright ownership.
- Copyright management: Blockchain enables content owners to directly manage their works.

The MediaVerse copyright management tool provides a machine-readable format for content creators to handle the legal aspects of copyright. The platform also provides a legal framework to allow storage and registration of assets and smart negotiation of (multimedia) content to manage revenues.

5.1 Testing MediaVerse Copyright Management Tool

MediaVerse aims to validate its copyright management framework among three major use cases: Citizen Journalism; Co-creation of immersive and inclusive media; An artistic experiment under the headline of “Hybrid Intelligence”. This paper focuses on the use of the blockchain technology linked to the second use case “Co-creation of immersive and inclusive media”. More specifically, it is linked to the findings gathered from two focus groups conducted with professionals working in Audiovisual Translation and Media Accessibility fields. The aim of the research was twofold. First, to gain information about audiovisual translation needs and expectations of the MediaVerse platform in relation to rights management of media accessibility assets from a user-centric approach. Second, gather and analyse data from users to understand the existing workflow for production, distribution, and monetisation of digital assets in their fields.

Methodology

This section describes the methodological steps and tools used during the focus groups. In both cases, participants were recruited via e-mail. The choice of recruitment was having experience in the AVT and/or MA fields at a professional level. The focus group duration was agreed with participants as 60 minutes and was held online using Microsoft Teams (Teams, 2022) video conference platform.

The selected methodological tools to gather quantitative and qualitative data from participants were an online questionnaire and a focus group. At the beginning of the session, ethical procedures were strictly followed to ensure compliance with EU existing regulations and codes of conduct. Second, a demographic questionnaire was provided to participants. Third, the focus group was conducted, (details are provided under section 5.1.3). Finally, a satisfaction questionnaire related to the MediaVerse platform was provided to participants. Conclusions gathered from the satisfaction questionnaire and the focus group were validated with all participants.

Demographics

Following the COnsolidated criteria for REporting Qualitative research (COREQ) based on Tong et al. [8] we report the criteria used for demographics. The interviewer and facilitators are the authors of this paper, all females with PhD. In terms of demographics the profile of the participants were nine females and one male. All participants were actively working in different modalities of the AVT and MA fields (Live subtitling, AD, Subtitling for the scenic arts, SDH, AVT), and 5 of them with active teaching duties at BA and MA level in Translation Studies for different Spanish higher education institutions. Six of them hold a PhD and four a MA. Five reported more than 10 years of experience in the AVT and MA fields, two reported 6-10 years, and four reported 3-5 years of experience. All participants reported that their main language combination of work was English into Spanish, and four of them also included English into Catalan. Other reported language combinations were Italian, German, and French into Spanish/Catalan. All participants reported that their work was based in Spain.

Focus Group Procedure

After a short introduction, a theoretical presentation of blockchain technology as part of the MediaVerse platform to manage copyrights was explained through a set of slides in a PowerPoint Presentation. Second, the list of the possible blockchain-based solutions was also presented. Third, a discussion among participants was conducted by the facilitator and a designated note taker was responsible for taking notes. The discussion was structured around the following three questions:

1. Do you think that the MediaVerse platform could be used in the audiovisual translation field? Does it have any advantage compared to the current way of managing the different modalities of the AVT field?
2. Within the frame of accessibility and audiovisual translation files (i.e, media accessibility assets) rights management, authors have the moral right over the assets

they create. This can never be sold. Thus, assets should be somehow minted for moral ownership. Do you agree?

3. Should authors be able to establish the economic rights and rights of exploitation?

At the end of the focus group, extracted conclusions were gathered and validated by all participants. Participants were also invited to answer an evaluation questionnaire in relation to: the use of MediaVerse in their professional and teaching contexts, and the level of relevance of the proposed possible blockchain-based solutions in the AVT and MA fields.

Extracted Conclusions and Results

The replies to the three questions are outlined below, along with the results of the satisfaction questionnaire.

Question 1: Do you think that the MediaVerse platform could be used in the audiovisual translation field? Does it have any advantage compared to the current way of managing the different modalities of the AVT field?

All participants consider that the MediaVerse platform could be useful in the professional domain of audiovisual translation and media accessibility, although its use will depend on the type of content. While in the AVT field, most participants manage their copyrights through DAMA, in the MA field copyrights remain in an uncertain area and are not subjected to copyrights. One of the main problems in both cases is that authors cannot trace what use is made of their work (i.e. subtitles, audio description, etc.) once they deliver it, for instance in another country/territory). In this regard, the platform would be useful to track the further use of their work in other countries/territories or even in other platforms. In the specific case of film festivals, most of the time AVT professionals receive support materials (i.e. subtitling templates) but do not know its origin (i.e. author). In addition, once they deliver their work (i.e. subtitles), they are unaware of its possible further use. All participants state that they work in the context of Spain, with the Spanish language as their target language, so they are unaware about the use of their work in other countries/territories or platforms. One participant states that the agencies/clients she works with manage copyrights themselves, sometimes through specific platforms.

In the specific case of MA one participant points out that in live subtitling, the work is done in the abstract, and there are various people involved in the subtitling process, making it difficult to assign copyright to specific individuals for this specific modality. The participant asserts that translation agencies keep the exploitation rights. In live subtitling, when programs are long, subtitling is done by several subtitlers because they have to take turns. Likewise, in contexts such as television, the company that provides the subtitling service is often external to the media, so the question arises: to whom do the rights belong, to the company that provides the services or to the television that commissions the work? Finally, a participant points out that the platform would be useful for sharing audio descriptions, as copyright in this modality is in a gray area.

Question 2: Within the frame of accessibility and audiovisual translation files (i.e., media accessibility assets) rights management, authors have the moral right over the assets they create. This can never be sold. Thus, assets should be somehow minted for moral ownership. Do you agree?

All participants agree on the preservation of moral property rights. This is a particularly important issue for participants who work with modalities derived from media accessibility services, since “services such as audio description or subtitles for the deaf and hard of hearing are not considered to be original work worth right protection according to copyright laws” (Orero et. al 2023:10).

Participants also note that there has been a lot of progress in recognizing moral rights in recent years, thanks to the efforts of professional associations in the industry. One participant brings up the issue of automatic translation and moral ownership, asking how rights are recognized in cases of automatic subtitling and who owns the work in such situations. This question is becoming a concern and a priority in the European Agenda, also due to the increasing use of Artificial Intelligence (AI) tools, such as Chat GPT, and its effect on copyright rules³.

Question 3: Should authors be able to establish the economic rights and rights of exploitation?

All participants consider that this is a complex issue, because in many cases economic rights and conditions are reflected in previous signed agreements. However, all agree that authors should be able to participate in the negotiations to establish the conditions of exploitation of their works.

One participant points out that live subtitling is very ephemeral, but rights should be exploited, since many times the subtitles generated live are edited or used for later reruns (mainly in the television context). In the case of recorded conferences, the initial subtitles are also later edited, but exploitation rights are not established.

Another participant mentions that there is an increase in the translation of audio descriptions, and it is necessary for professionals in this field to register the corresponding rights. In this regard, another participant emphasizes the importance of professionals' involvement in the negotiation processes for the management of exploitation/distribution rights to prevent this accessibility service from being centralized solely through an entity (e.g., ONCE, a Spanish public-law corporation created to provide services for people with visual disabilities)⁴ and restricted to users of this organization.

In the case of audio descriptions and subtitling for performing arts, this recognition may have a short life span, since productions are punctual/limited and subject to modifications.

In the evaluation questionnaire, participants were asked to share their opinions on the advantages and disadvantages of the MediaVerse platform and which blockchain solutions they thought would be most relevant for professional translators.

³ https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/intellectual-property-chatgpt-2023-02-20_en

⁴ <https://www.once.es/>

Participants reported the following advantages: potential use for copyright recognition and distribution of accessible content, an intuitive platform, ease of sharing content, and the ability to track assets. However, they also pointed out some disadvantages, such as the fact that copyright is sometimes a gray area in the translation field, the early development stage of the platform, the challenge of recruiting a broad group of users, and the complexity of professional relationships with companies in the field.

Regarding the proposed set of blockchain-based solutions, participants reported that the most relevant were a decentralized digital content ecosystem, attribution, and copyright management.

Limitations of the Study

The findings of this study have to be seen in light of some potential limitations that could be addressed in future research. First, the study is focused on the Spanish context, considering only AVT and MA professionals working in this country. Including a wider number of countries could prove to be important not only to validate the proposed MediaVerse blockchain-solutions for copyright management in the AVT and MA fields, but also to improve an exchange of good practice among professionals in both fields. The second limitation could be considered in terms of gender bias, as nine out of ten participants recognized themselves as females. It could be argued that contrary to other fields of the audiovisual industry, professionals in the AVT and MA fields are mainly females. Still, future research related to copyright management should include a broader gender representation in the sample of participants, not only from a binary perspective (i.e. male and female), but also in terms of non-binary participants.

6 Conclusions

To effectively foster diversity, pluralism, and democratic values within the media industry and the constant evolution of the Web, a new approach to content management is required. In the context of Web 3.0, MediaVerse could provide a standardized framework for managing ownership and licensing of digital assets, which can help to promote legal and ethical sharing and reuse of digital assets in a decentralized and open environment. The use of blockchain ensures that all transactions and ownership records are tamper-proof and cannot be altered.

To sum up, as it has been explained along the article, blockchain technology can transform the way in which media content creators manage and share the intellectual property rights of their digital assets. This technology has been included as part of the MediaVerse platform in order to allow content owners assigning copyrights and tracking the data they generate in a secure and transparent network, such as the transactions/interactions between peers. Professionals in the AVT field are content creators, and their works (i.e. subtitles) are considered derivative works, therefore authors of these digital assets have the right to be protected under copyright. On the other side, the work of professionals in the MA field remain in an uneven field as their work is not equally protected in all countries. For instance, while in France AD are subjected to copyrights, this is not the case in Spain.

With the aim of testing the suitability of the blockchain based solutions proposed under the MediaVerse platform, a focus group with professionals from the AVT and MA field was held. As reported in the extracted conclusions, participants agree on the potential use of the platform for copyright recognition, attribution and distribution of accessible content, and describe it as an intuitive platform, where creators and consumers can easily share content and can track their digital assets.

Acknowledgments

This article is part of the MediaVerse project (Grant agreement 957252). One of the authors is part of TransMedia Catalonia research group, funded by the Catalan Government through the SGR funding scheme (Catalan Government funds, 2021SGR00077).

References

1. Bonnet, S., Teuteberg, F.: Impact of blockchain and distributed ledger technology for the management, protection, enforcement and monetization of intellectual property: a systematic literature review. *Information Systems and e-Business Management*. 10.1007/s10257-022-00579-y. (2022).
2. Geradin, D. (2021): What is a digital gatekeeper? Which platforms should be captured by the EC proposal for a digital market act?' Tilburg Law and Economics Center (TILEC); Geradin Partners; University of East Anglia (UEA) — Centre for Competition Policy; University College London. <http://dx.doi.org/10.2139/ssrn.3788152>.
3. Levis, D., Fontana, F., Ughetto, E.: A look into the future of blockchain technology. *PLOS ONE* 16(11) (2021): e0258995. <https://doi.org/10.1371/journal.pone.0258995>.
4. Matamala, A., Soler-Vilageliu, O.: Defining and assessing artistic co-creation: the TRACTION proposal. *Arte, Individuo y Sociedad*; 34(3): 851–867 (2022).
5. Oncins, E.: From stage to screen: Digital transformations and accessibility in the scenic arts. Kostopolou, L. and Misiou, V. (eds.). *New Paths in Theatre Translation and Surtitling*. London: Routledge. (forthcoming). Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999).
6. Orero, P., Fernandez Torner A., Oncins E.: The visible subtitle: Blockchain technology towards right management and minting. *Open Research Europe* 2023, 3:26 (2023).
7. Sánchez García, R.: La propiedad intelectual en la España contemporánea, 1847-1936. *Hispania*, 62(212), 993–1019 (2002). doi.org/10.3989/hispania.2002.v62.i212.247
8. Tong King, L.: Translation and copyright: towards a distributed view of originality and authorship. *The Translator*, 26:3, 241-256 (2020). [doi: 10.1080/13556509.2020.1836770](https://doi.org/10.1080/13556509.2020.1836770)
9. Tong A, Sainsbury P, Craig J.: Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care*; 19(6): 349–357 (2007).
10. Wensheng, G., Zhenqiang, Y., Wan, S., Philip, Y. (2023): Web 3.0: The Future of Internet. Companion Proceedings of the ACM Web Conference, 2023, pp. 1–10 [doi: 10.1145/3543873.3587583](https://doi.org/10.1145/3543873.3587583).

Developing a Customisable Subtitling Tool Based on Academic Research and User Needs

Željko Radić [0009-0003-9451-5557]
University of Surrey, Guildford, Surrey GU2 7XH, UK
z.radic@surrey.ac.uk

Abstract. SpeakSubz is a tool for translating non-live subtitles primarily relying on Google's speech recognition technology: Google Voice Typing. This specialised software also includes functionalities that make it a viable tool for academic research on various topics related to subtitling and voicing/respeaking. It was used as a tool within a larger doctoral study called 'Interlingual Subtitle Voicing: A New Technique for the Creation of Interlingual Subtitles, A Case Study in Croatian'. This study aimed to measure speech recognition accuracy for Croatian subtitles, the duration of various elements of a subtitling workflow, and, ultimately, to assess participants' reception of the technique and the tool and elicit their feedback for its future development. The tool is being developed based on academic scholarship, industry insights and, most importantly, subtitlers' needs. Within the framework of action research and the notion of subtitler experience and as highlighted in this paper, the tool is developed in cycles akin to software updates. In each cycle, a major new variable is introduced and tested with other minor changes related to either functionality or the interface. This paper will also highlight some of the most relevant functionalities that distinguish the current version of SpeakSubz from similar tools. The goal is to create a customisable tool for human-informed translation of subtitles that can be used for training and research in academia as well as professional environments.

Keywords: SpeakSubz, ISV, Subtitling, SUBX, Action Research, Software Development, Research Cycles

1 Introduction

Interlingual Subtitle Voicing (ISV) is a technique devised for translating non-live subtitles by voicing, traditionally known as respeaking. SpeakSubz, a specialised tool, was created to assess the technique's viability. This tool enabled quality measurements of target language output, emphasising speech recognition (SR) accuracy, as well as durations of various subtitling workflow elements such as typing, voicing and editing. Eventually, the ISV study resulted in subtitlers' evaluation of both the technique and the accompanying software, which was crucial for further development.

This strand of research started with an MA thesis completed at the University of Roehampton in 2018 [1], which resulted in the desktop version of the tool (the first ISV cycle) from an improvised methodological setting: uploading subtitles to the Google Voice Typing (GVT) interface and translating them by voice. The desktop version of SpeakSubz was further developed as part of a doctoral thesis currently being finalised at the University of Surrey. After internal testing within the Centre for Translation Studies (CTS) at Surrey and the pilot study with students, an online version of SpeakSubz was created (the second ISV cycle), further improved after the main experiment.

2 Development Cycles

Per action research [1]–[3], the ISV doctoral study was exploratory, participatory and conducted in cycles. Each cycle was informed by relevant academic research, industry experiences at that time and feedback coming from participants of pilot and main studies. This enabled the shaping of the tool according to future users' needs and preferences, in line with the concept of agile software development [5]; it also offered them multiple functionalities that they could use as needed, based on individual preferences. The tool was developed by a Croatian software developer, A. Vrečić, and integrated with the ISV website¹ created by a Croatian web designer, A. Prskalo.

The ISV website, primarily the result of the Covid-19 pandemic, is an online hub for experiments containing all the necessary information and materials for ISV training and testing. Such interdisciplinary cooperation enabled a streamlined training and testing environment for the new technique. The following sections will chronologically describe the cycles of the software development process in brief.

2.1 Desktop SpeakSubz

As mentioned above, the desktop version of SpeakSubz resulted from an MA study that compared the speeds of typing subtitles interlingually versus voicing them. The experiment was conducted in the GVT application [6] in Google Docs. Participants had to copy subtitles into the GVT interface and translate them by voicing. At the same time, they had to track and note down the times they spent on each exercise, including possible breaks. In addition, they had to also record their process with an external tool: Screencast-o-Matic (now ScreenPal²).

¹ <https://isvresearch.eu/>

² <https://screenpal.com/>

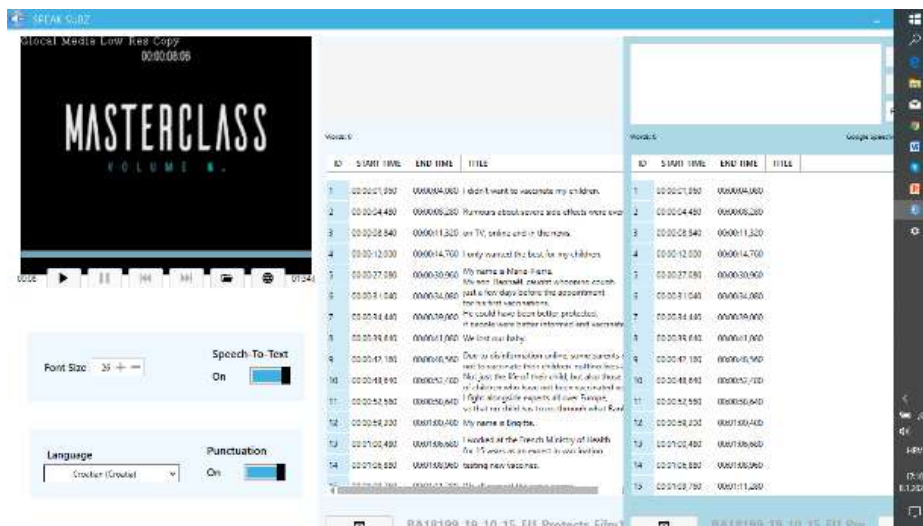


Fig. 1. Desktop SpeakSubz Interface

Besides collecting empirical data, participants were also invited to give feedback about the technique, which, among other factors, shaped the next iteration of desktop SpeakSubz (Figure 1). This version was also used for internal testing within CTS at Surrey and piloting with translation students from the Faculty of Humanities and Social Sciences at the University of J.J. Strossmayer in Osijek, Croatia. Although some of the piloting was initially supposed to occur face-to-face, because of the Covid-19 pandemic, the methodological process had to be transferred online. This resulted in the ISV website that contains all the information about the ISV technique and the software in the form of PowerPoint and video presentations, as well as short video tutorials on various functionalities of SpeakSubz. This piloting process resulted in an online SpeakSubz version integrated with the ISV website.

2.2 Online SpeakSubz

The online iteration of SpeakSubz was a more advanced version because it included a plethora of functionalities suggested by participants of the previous experiment cycles. In Figure 2, for example, in comparison with the desktop version of SpeakSubz subtitle boxes are much bigger, there is a big green button to activate or stop a speech recognition session, visual reading speed markers (thin green bars that turn red once the reading speed is exceeded) and comment boxes. The interface is similar to some existing subtitling platforms, especially cloud-based ones, to evoke a sense of familiarity.

Online SpeakSubz was created in dark mode, per software development industry standards. Certain academic research also explores various aspects of the dark mode. Some studies show that dark mode reduces eyestrain in lower ambient illumination [7] and increases visual acuity, which is important for an efficient and pleasant subtitler experience (SUBX). In addition, one study shows that dark mode

could potentially save energy [8]. This is especially relevant for SpeakSubz because it is aimed to be used on mobile devices such as tablets or hybrid laptops and away from energy outlets, thus becoming an ergonomic tool.

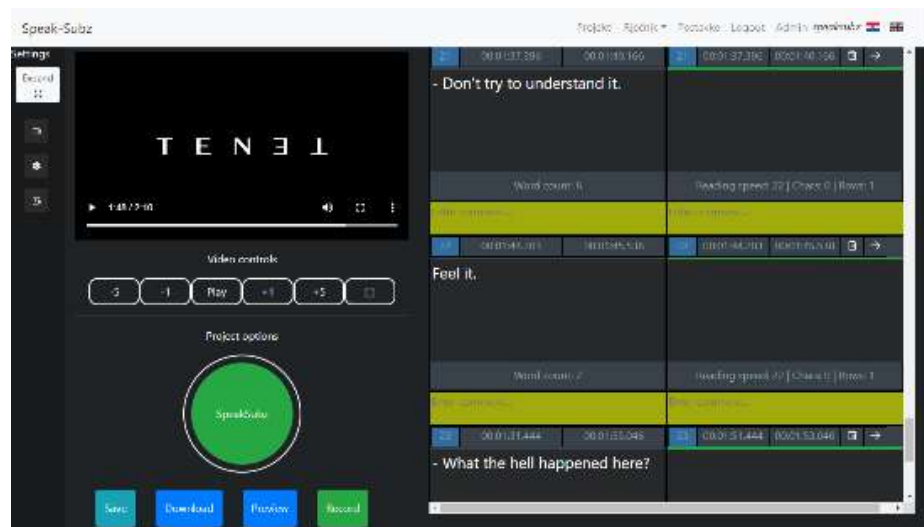


Fig. 2. Online SpeakSubz Interface

SpeakSubz is an example of how a tool can improve the existing underlying technology and tailor it to specific needs. For instance, the desktop version of SpeakSubz introduced voicing punctuation, which even GVT does not offer for Croatian. The online version went even further and enabled entering SR misrecognitions into the SpeakSubz dictionary for better accuracy. The online version of SpeakSubz also includes automatic spell-checking to increase output quality further. This version of SpeakSubz also shows a number of characters per row, which is critical for readability in non-live subtitling.

Regarding research methodology, the online version of SpeakSubz also has an integrated screen and voice recording functionality and pre-loaded tasks that had to be loaded and exported manually in previous versions. This automation of various processes resulted in participants' positive experiences with the ISV technique and SpeakSubz software, resulting in positive reviews – scoring above 4 out of 5 (4.06 and 4.13, respectively) – despite imperfections in the underlying technology (GVT). This data stems from post-testing questionnaires (15 participants) and optional interviews (6 participants) conducted after the main ISV experiment, providing us with a wealth of information crucial for future developments.

To sum up, SpeakSubz is a tool that aims to combine different functionalities needed to create non-live subtitles in multiple hybrid workflows. The interface's similarities to existing online solutions help users adjust to the tool. However, most current subtitling software does not include SR as an input method, even for major languages, leaving much space for exploration. Additionally, SpeakSubz enables entering misrecognitions, which the underlying technology, GVT, does not offer in its interface. This makes

further improvements possible in SR accuracy, especially for lower-resourced languages.

Regarding teaching and research, SpeakSubz's integrated recording functionality allows users to re-watch their performance, reflect on it and improve aspects that need to be practised. Professional subtitling software generally used in audiovisual translation (AVT) research usually does not offer such functionality. Researchers, in addition, can utilise these recordings to observe the translation process and participants' behaviours. In short, SpeakSubz offers functionalities that facilitate methodological processes that most professional desktop or cloud-based software like WinCaps, Spot, EZTitles and Oona do not provide, primarily to protect copyrighted source materials.

Based on the above and to my knowledge SpeakSubz would be the only subtitling tool from academia tailored to academic teaching, training and testing that offers technologies and functionalities – such as SR input and integrated recording – that professional tools do not provide. Furthermore, SpeakSubz is highly adaptable and can include and adapt new and emerging technologies in each new cycle to further test them to gauge whether they increase efficiency. Such an improved tool could then be presented to industry stakeholders, both subtitlers and clients.

3 Future Development

Based on the data from the main ISV experiment, the following goals have been established for the subsequent research and development cycle (the third ISV cycle): introducing a new major technology into the workflow (machine translation), testing the technique and the software on mobile and touch-screen devices and adding other functionalities suggested by participants in the ISV main experiment from the second cycle. For example, the current version of SpeakSubz includes virtual buttons for adding punctuation by touch as an alternative to voicing or typing them; this feature is yet to be tested and evaluated within the ISV workflow.

These elements will be included in the next stage of testing hybrid workflows. However, this time, we will provide the participants with various functionalities and measure and test what they prefer the most and why so that the tool can be customised even more in future. These will be explored under the concept of subtitler experience (SUBX) introduced in the ISV study, stemming from translator experience – (TX) [9–11], the term that Zapata derived from the general user experience (UX). In short, SpeakSubz will be explored not only as a training and testing, as well as a professional tool but also as an ergonomic one that enables subtitlers to work in more comfortable environments, away from the traditional keyboard-desk setting.

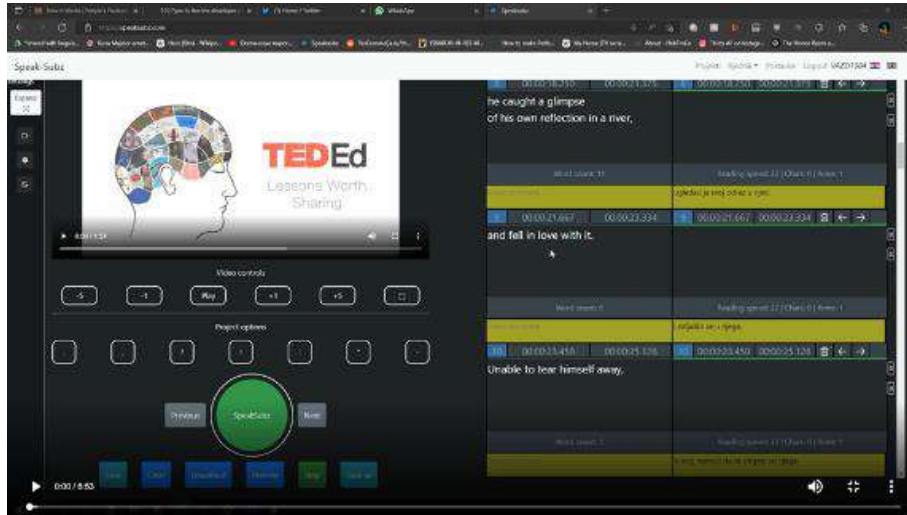


Fig. 3. Virtual Punctuation Keys in Online SpeakSubz

4 Conclusion

This demo paper describes the development of specialised software based on academic research for academia, primarily for creating and quantitatively testing hybrid subtitling workflows and new technologies. The software also enables gathering qualitative data about the specialised subtitling tool and the new technique being developed around it. Suppose empirical research shows this tool can increase the efficiency of subtitlers while keeping in mind their physical and mental well-being. In that case, it might also be developed commercially and offered to major clients in the AVT industry.

This interdisciplinary and participatory research not only gathers data from participants but also teaches them new skills and aims to inform them about the benefits of ergonomics, which can be done anonymously online. Such an environment is conducive to collecting objective data and ample qualitative feedback, with one main goal: to research what subtitlers need and how this novel form of human-machine interaction can produce the best results.

References

1. Radić, Ž.: Speech Recognition in Interlingual Subtitling: A Case Study in Croatian, (2018).
2. Cravo, A., Neves, J.: Action research in translation studies. *Journal of Specialised Translation*. (2007).
3. Massey, G.: Applied translation studies and transdisciplinary action research: Understanding, learning and transforming translation in professional contexts. *Russian Journal of Linguistics*. 25, (2021). <https://doi.org/10.22363/2687-0088-2021-25-2-443-461>.
4. Neves, J.: Action research: So much to account for. Presented at the (2018). <https://doi.org/10.1075/bct.95.05nev>.
5. Dingsøy, T., Nerur, S., Balijepally, V., Moe, N.B.: A decade of agile methodologies: Towards explaining agile software development, (2012). <https://doi.org/10.1016/j.jss.2012.02.033>.
6. Google Docs Editors Help: Type with your voice.
7. Xie, X., Song, F., Liu, Y., Wang, S., Yu, D.: Study on the effects of display color mode and luminance contrast on visual fatigue. *IEEE Access*. 9, (2021). <https://doi.org/10.1109/ACCESS.2021.3061770>.
8. Dash, P., Hu, Y.C.: How much battery does dark mode save?: An accurate OLED display power profiler for modern smartphones. In: *MobiSys 2021 - Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (2021)*. <https://doi.org/10.1145/3458864.3467682>.
9. Zapata, J.: Exploring multimodality for translator-computer interaction. In: *ICMI 2014 - Proceedings of the 2014 International Conference on Multimodal Interaction (2014)*. <https://doi.org/10.1145/2663204.2666280>.
10. Zapata, J.: Translators in the Loop: Observing and Analyzing the Translator Experience with Multimodal Interfaces for Interactive Translation Dictation Environment Design, (2016).
11. Zapata, J.: Translating On the Go? Investigating the Potential of Multimodal Mobile Devices for Interactive Translation Dictation. *Tradumàtica: tecnologies de la traducció*. (2016). <https://doi.org/10.5565/rev/tradumatica.180>.

A pedagogical platform for spoken post-editing (PE): the integration of speech input into COPECO

Jeevanthi Liyana Pathirana¹, Pierrette Bouillon² and Jonathan Mutal³

^{1,2,3} Faculty of Translation and Interpreting, University of Geneva
juliyanapathirana@gmail.com, Pierrette.Bouillon@unige.ch,
Jonathan.Mutal@unige.ch

Abstract. We present the integration of speech input into COPECO, an online teaching and learning platform developed to collect learner translation/post-editing corpora[22]. Speech offers potential to increase translators' productivity and wellbeing by reducing typing time and effort. A multimodal, speech-enabled COPECO will complement other researchers' and developers' efforts to integrate speech dictation and post-editing into translation tools, training and practice (e.g. MateCat allows post-editors to activate a speech-to-text component or TradDICT Learn[28], an online-learning platform to develop sight-translation and dictation skills). COPECO helps trainers to compare different translation modalities, prepare courses and share teaching resources. To our knowledge, this is the first online platform to allow comparison between written and spoken post-editing and gathering data/statistics on speech-based post-editing behaviour (e.g. respeaking whole segments vs. correcting minor errors, etc). This helps translation trainers understand frequent errors made by learners when using speech which can then be used to improve their course content. It will also help to build a "speech post-editing corpus" which will be beneficial in the long run for research and analytical purposes. Unlike proprietary speech recognition add-ons with defined services, we develop speech commands for post-editing based on translators' behaviour and needs, optimising COPECO based on requirements with minimal costs.

Keywords: Automatic Speech Recognition (ASR), Post editing (PE), Computer Assisted Translation (CAT)

1 Introduction

Translation services are used for translating documents into multiple languages, within a limited time frame, with high accuracy. Post-editing of Machine Translation (MT) is known to allow translating large volumes of translations while saving costs and time [29]. Workflows in the translation industry have experienced a significant transformation in a way that speech technology is likely to contribute to further innovation [8]. Preliminary studies on speech based post-editing [14][15][16] show that provided that Automatic Speech Recognition (ASR) and MT output are of high quality and that the translators are competent with software (Computer Assisted Transla-

tion (CAT) tools, MT suggestions and ASR toolkits such as Dragon [9]), speech based post-editing can be a promising approach which can result in performance gains in the translation workflow. However, for speech based post-editing to be used by translators, this modality has to be properly introduced to them. For that, translation trainers should analyze behaviors of speech based post-editing and translation, to then use that knowledge to improve their courses on translation training.

In this paper we demonstrate COPECO [22], a speech-enabled platform developed for collecting learner translation/post-editing corpora and for helping translation teachers annotate and learn student post-editing behavior. First we give an overview of speech based translation/post-editing in previous studies, leading towards the rationale behind our work and then we demonstrate the design of our setup and its functionalities.

2 Speech-based PE

Computer Assisted Translation tools are mostly based on traditional input such as keyboard and mouse [27]. However, the translation industry constantly seeks ways to improve speed and quality by incorporating various technological advancements into translation workflows [3][21][24].

One such technology is Automatic Speech Recognition (ASR) which automatically transcribes spoken input into text[7]. As a result, a large body of literature has explored the integration of ASR into translation processes and the ways in which speech technologies can be effectively utilized in translation[8][10][20][25][27][32][34]. Previous studies described successful use of ASR by freelance translators, with more productivity, allowing for "more flexible, translator-centered, ergonomic workflows and workspaces"[6][7]. Commercially available CAT tools have started offering integration with ASR systems as well, aiming for increased productivity and ergonomics during the translation process. Some examples are memoQ [19] combined with Apple's speech recognition service or Matecat combined with the ability to dictate the translation[17][18]. With some other CAT tools [30], commercial ASR systems for dictation, such as Dragon Naturally Speaking[9] can be used.

Recent studies researched and surveyed [20] the potential of using ASR for post-editing purposes. A study investigating the effects on productivity and on a translator's experience of integrating machine translation post-editing with speech technologies revealed that post-editing with the aid of a speech recognition system was faster than translating with the aid of a speech recognition system and also less tiresome (i.e., more ergonomic)[34]. Similarly, studies that looked into the possibility of using speech technologies for post-editing purposes in international organizations revealed translators were open to try speech-based post-editing as a new translation workflow [14]. Another study[11] found voice input more interesting than typing only when post-editing, as when some segments require major changes they could be dictated. If the post-editor is not a touch-typist, it was also found that the back and forth transfer of visual attention between source text, machine translation output and keyboard adds to the complexity of the task.

Many studies investigated the possibility of having multimodal interfaces to improve translation workflows, where ASR was a component. One study observes and analyzes translator experience (TX) with off-the-shelf voice-and-touch-enabled multimodal interfaces, as opposed to the interaction with traditional keyboard-and-mouse graphical user interfaces, to provide better recommendations for translation tool design [33]. Later, [27] developed a web-based translation editing interface that permits multimodal input via touch-enabled screens and speech recognition in addition to keyboard and mouse, which demonstrated the importance of the ASR quality rather than the features of the interface. [12] presented and evaluated the MMPE CAT environment that explores the use of speech commands, handwriting input, touch reordering, and multi-modal combinations for PE of MT. MMPE was later improved with additional speech and other facilities in the next version [13] along with some correction commands based on speech input. Unlike [27] which used a web-based interface, this environment requires specific hardware, which obliges the users to have this configuration in their translation environment.

These studies confirm that speech as an input modality does indeed provide promising improvements in the translation workflow, which supports it being included in the translator training process. Previous studies[2] mention current occurrences of audiovisual translation (AVT) modules incorporated in translator training programs, where use of ASR and TTS in AVT is specially studied in specialist courses. The study also emphasizes the importance of translator trainers being acquainted with the professional environment and the latest trends. There already exist courses such as Tradict Learn [28] which provides introductory courses to learn interactive translation (using voice) as well.

However, if we want to understand the frequent errors and behaviors of these modalities, it is necessary to analyze and understand translator data and statistics. Today, many tools exist for monitoring written-based PE (e.g. PET [1]), but when it comes to analyzing speech based inputs for translator training, there are only a few examples. Workbenches like Matecat [17] enable speech based input for translating via dictation [18], but do not contain enough data protection or the ability to revise using customized annotation schemes. A recent study experiments with the combination of speech synthesis and PE where they investigate the benefits and drawbacks of exposing students to novel technologies and practices such as synthetic voices and PE early, and the role these interactions can play in translator training [4]. However, very few have considered using data for inferring translator training techniques on ASR based techniques.

3 COPECO Design

Most previous studies on speech based post-editing and translation have been based on tools that require specific hardware requirements based standalone applications and other studies require commercial license based software such as Dragon[8] and Trados Studio workbench [30]. For our work, we chose the tailor-made open source PE platform COPECO [22] to integrate speech recognition. COPECO was originally

developed as a project to collect post-edits produced by students and teacher corrections and to structure the task of translation error annotation. The aim of COPECO is to translation teachers with an online post-editing platform, designed to help them to annotate student post-editing tasks using a shared or personalized annotation scheme. In the long run, data collected from a platform like COPECO can also be used to analyze how post-editing tools can impact the translation industry economy.

These features made COPECO a suitable platform to integrate speech so that speech based translation or post-editing can be done. Currently, speech integrated COPECO web based platform allows the translation trainers to assign tasks with text to students to translate and their machine translation suggestions. The student can then translate from scratch or post-edit the translation, using either typing, speech, or a mix of both. Once the task is translated, the student can submit the task to the teacher. The teacher can then correct the task using systematic translation error annotation [23]. Previous studies have worked on defining frameworks on error analysis of MT [31] or ASR[26] outputs. COPECO allows error annotation with predefined translation schemes or their custom annotation schemes. COPECO also allows to visualize the corpus with the translations, corrections, reference translation (if it exists) as well as the annotations. It simultaneously builds an open source student post-editing corpus by collecting post-edits produced by student and teacher corrections. All data are collected and can be anonymously shared.

For speech integration, we used a publicly available ASR engine powered by Google Web Speech, which connected to our tool via an application programming interface (API)[5]. The Web speech API only allows speech recognition, and no commands were available for post-editing. So an initial set of speech commands used for post-editing were developed. Some of the commands developed for English include "clear segment", moving the cursor to specific locations, selecting a word/phrase and replacing it, deleting words/phrases, saving a segment and moving to the next segment. Table 01 shows a subset of the speech commands that works for English language. Similarly, speech commands can be developed for any other target language that we would need to use speech based translation and post-editing tasks.

Table 1. Example set of speech commands for post-editing in English

Command	Description
"ABC"	Inserts ABC in the current cursor location.
"Select XYZ"	Highlights XYZ (can be a word or a phrase).
"Undo Highlight"	Removes the highlighting of XYZ.
"Delete that"	Deletes the highlighted word/phrase.
"Begin line"	Moves cursor to beginning of segment.
"Finish line"	Moves cursor to end of segment.
"Next/Previous Segment"	Saves the current segment and then moves to the next/previous segment.
"Save segment"	Saves the current segment.
"Clear segment"	Clears text in the current segment.
"Comma" "Period" "Question mark" "Semicolon" etc.	Inserts punctuation marks ",", ".", "?", ";" in the current cursor location.

Figure 01 shows when the student has opened the PE task, where the source language is French and the target language is English. The student can speak (by clicking the microphone image), type or speak commands for each segment. In a PE task, the MT suggestions that were included when the task was created will appear on the target (right) side. In a translation task, the target side will be empty by default (the MT suggestion can however be made visible below the source segment by clicking a button, if needed.)

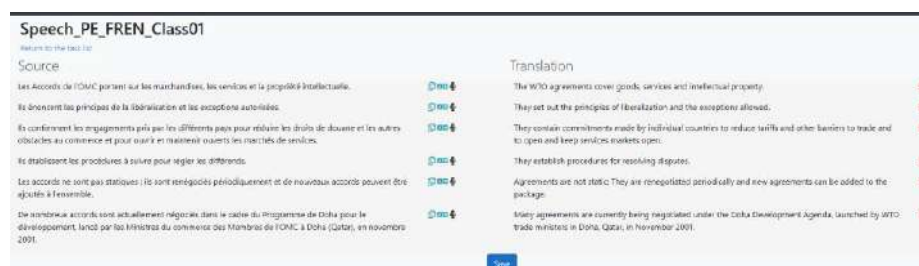


Fig. 1. Student view of a task assigned by the teacher

Figure 02 shows when a student uses spoken post-editing in the second segment. The student has switched the microphone on and had asked via a speech command to select the phrase “exceptions allowed”, which is highlighted. The speech command is printed below, under Speech Commands, with a time log. Once a segment is selected, the student can either delete, replace or undo the selection.

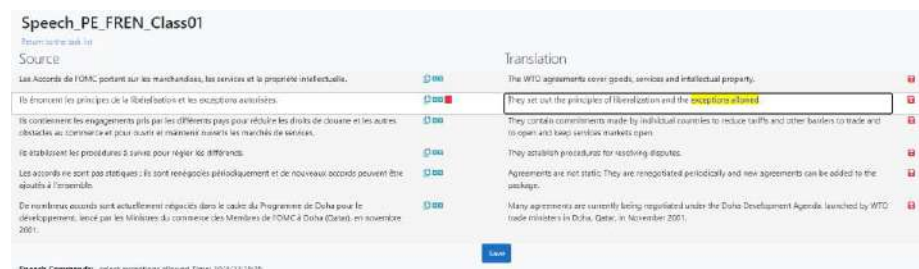


Fig. 2. Using spoken commands to post-edit the segment (e.g. "select exceptions allowed")

As shown in Table 01, some other spoken commands can be used to move in between segments, save segments, clear the entire segment, move to beginning and end of the segment and to speak out punctuation marks. Some other sub commands perform capitalizing the beginning of the sentences/specific words and phrases. All these commands can be developed for other target languages as well upon need.

Once the translations are complete, the student submits the task. The submitted tasks will be sent to the corrector/teacher, along with the statistics of each segment: the number of keystrokes, time taken and average statistics of each task as well. The corrector will then annotate the errors using pre-defined/pre-imported error annotation

schemas. Figure 03 shows the corrector annotating a terminology error in one of the segments.



Fig. 3. Corrector annotates a terminology error found in the translation

Once the correction task is complete, detailed information and statistics are provided on the type and description of corrections made. Figure 04 displays one such log, where it shows the reference, source, machine translation target, postedit and the corrections made by the teacher, along with the type of error annotation. These statistics on both translator and corrector side, along with the corpora, allow translation trainers to better understand spoken translation/post-editing behavior.

Errors				
Type	Count	Description	When?	When?
Spelling	1	Words related to the form or content of a text (misspellings or whether it is a translation or not)		1
Terminology	1	A word (noun-verb) used in a particular way or with a different meaning than the one expected for the context or situation (lexical)		1
Number counting	1	Words related to counting of items		1
Verb conjugation	1	The text is incomplete		1

Production									
Ref	Source	Target	ICTarget	ICTargetCorrected	Scoreline	Targetline	Keystroke	Additional time	
The WTO agreements cover goods, services and intellectual property.	Les Accords de l'OMC couvrent les marchandises, les services et la propriété intellectuelle.	The WTO agreements cover goods, services and intellectual property.	The WTO agreements cover goods, services and intellectual property.	The WTO agreements cover goods, services and intellectual property.	0.00	0.00	0	0.00	
They set out the principles of liberalization and the permitted exceptions.	Ils énoncent les principes de libéralisation et les exceptions autorisées.	They set out the principles of liberalization and the permitted exceptions.	They set out the principles of liberalization and the permitted exceptions.	They set out the principles of liberalization and the permitted exceptions.	214.62	0.00	16	0.00	
This clause includes individual commitments to lower tariffs and other barriers to trade and to open and improve services markets.	Le règlement des engagements pris par les différents pays pour réduire les droits de douane et les autres obstacles au commerce et pour ouvrir et améliorer certains services commerciaux.	This clause includes individual commitments to lower tariffs and other barriers to trade and to open and improve services markets.	This clause includes individual commitments to lower tariffs and other barriers to trade and to open and improve services markets.	This clause includes individual commitments to lower tariffs and other barriers to trade and to open and improve services markets.	60.91	0.00	22	0.00	
They set procedures for resolving disputes.	Ils établissent les procédures à suivre pour régler les différends.	They establish procedures for resolving disputes.	They establish procedures for resolving disputes.	They set procedures for resolving disputes.	12.04	0.00	4	0.00	
These agreements are not static. They are negotiated periodically and new agreements are added to the package.	Ces accords ne sont pas statiques. Ils sont négociés périodiquement et de nouveaux accords peuvent être ajoutés à l'ensemble.	Agreements are not static. They are negotiated periodically and new agreements are added to the package.	These agreements are not static. They are negotiated periodically and new agreements are added to the package.	These agreements are not static. They are negotiated periodically and new agreements are added to the package.	29.71	0.00	11	0.00	
Many commitments are currently being negotiated under the Doha Development Agenda, launched by WTO leaders in Doha, Qatar, in November 2001.	De nombreux accords commerciaux sont actuellement négociés dans le cadre de l'Agenda de Doha pour le développement lancé par les dirigeants du commerce des membres de l'OMC à Doha (Qatar), en novembre 2001.	Many commitments are currently being negotiated under the Doha Development Agenda, launched by WTO leaders in Doha, Qatar, in November 2001.	Many commitments are currently being negotiated under the Doha Development Agenda, launched by WTO leaders in Doha, Qatar, in November 2001.	Many commitments are currently being negotiated under the Doha Development Agenda, launched by WTO leaders in Doha, Qatar, in November 2001.	23.00	0.00	6	0.00	

Fig. 4. Detailed information and statistics on post-editing tasks and error annotation

4 Conclusion

In this paper, we present the different functionalities of COPECO, a speech input enabled online teaching and learning platform developed to collect learner translation/post-editing corpora as well as to learn speech based translation and post-editing behavior and statistics. Future work includes adding new custom voice commands when needed for multiple languages and improving the usability of speech modalities, based on user feedback. While currently we use Google Web Speech API, we can also integrate other ASR services in the future. This would make this platform also able to be used to compare post-editing behaviors (productivity, fatigue and engagement) based on different speech recognizers and different translator profiles.

References

1. Aziz, W., Castilho, S., & Specia, L. (2012, May). PET: a Tool for Post-editing and Assessing Machine Translation. In LREC (pp. 3982-3987).
2. Bolaños-García-Escribano, A., Díaz-Cintas, J., & Massidda, S. (2021). Latest advancements in audiovisual translation education. *The interpreter and translator trainer*, 15(1), 1-12.
3. Bowker L. (2002). *Computer-Aided Translation Technology: A Practical Introduction*. Ottawa, Canada: University of Ottawa Press.
4. Brockmann, J., Wiesinger, C., & Ciobanu, D. (2022, June). 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* (pp. 249-257).
5. Chrome Web Speech API. <https://developer.chrome.com/blog/voice-driven-web-apps-introduction-to-the-web-speech-api/> last accessed 2023/04/25
6. Ciobanu, D. (2014). Of dragons and speech recognition wizards and apprentices. *Tradumatica* 2014, 524–538. doi: 10.5565/rev/tradumatica.71
7. Ciobanu, D. (2016). Automatic Speech Recognition in the professional translation process. *Translation Spaces. A multidisciplinary, multimedia, and multilingual journal of translation*, 5(1), 124-144.
8. Ciobanu, D., and Secară, A. (2020). “Speech recognition and synthesis technologies in the translation workflow” in *The Routledge Handbook of Translation and Technology*. ed. M. O’Hagan (Milton Park: Routledge), 91–106.
9. Dragon Speech Recognition Solutions, <https://www.nuance.com/dragon.html>, last accessed 2023/04/25
10. Dymetman, M., Brousseau, J., Foster, G., Isabelle, P., Normandin, Y., and Plamondon, P. (1994). Towards an automatic dictation system for translators: the TransTalk project. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 1–4.
11. Garcia-Martinez, M., Singla, K., Tammewar, A., Mesa-Lao, B., Thakur, A., Anusuya, M. A., & Carl, M. (2014). SEECAT: Speech & Eye-tracking Enabled Computer Assisted Translation. In *European Association for Machine Translation: EAMT* (pp. 81-88).
12. Herbig, N., Düwel, T., Pal, S., Meladaki, K., Monshizadeh, M., Krüger, A., & van Genabith, J. (2020, July). MMPE: A multi-modal interface for post-editing machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1691-1702).
13. Herbig, N., Pal, S., Düwel, T., Shenoy, R., Krüger, A., & van Genabith, J. (2020, October). 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* (pp. 93-108).
14. Liyanapathirana, J., Bouillon, P., & Mesa-Lao, B. (2019, August). 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* (pp. 149-158).
15. Liyanapathirana, J., & Bouillon, P. (2021). Integrating post-editing with Dragon speech recognizer: a use case in an international organization. *Translating and the Computer* 43, 55-67.

16. Liyanapathirana, J. et al. Integrating Speech in Post-Editing (PE)-Comparison of two PE Interfaces. In: *New Trends in Translation and Technology (NeTTT)*. Rhodes Island, Greece. [s.l.] : [s.n.], 2022. p. 120–123.
17. MateCat Homepage, <https://site.matecat.com/>, last accessed 2023/04/25
18. MateCat Guide, <https://guides.matecat.com/translate-1>, last accessed 2023/04/25
19. memoQ Homepage, <https://www.memoq.com/>, last accessed 2023/04/25
20. Mesa-Lao, B. (2014, April). 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* (pp. 99-103).
21. Mossop B. (2006). Has computerization changed translation? *Meta* 51, 787–805. doi: 10.7202/014342ar
22. Mutal, J.D., Bouillon, P., Schumacher, P. and Gerlach, J., 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*.
23. O'Brien, S. (2011). Towards a dynamic quality evaluation model for translation. In *Journal of Specialized Translation*, 17:1–2.
24. O'Brien S. (2012). Translation as human–computer interaction. *Transl. Spaces* 1, 101–122. doi: 10.1075/ts.1.05obr
25. Reddy, A., and Rose, R. C. (2010). Integration of statistical models for dictation of document translations in a machine-aided human translation task. *IEEE Trans. Audio Speech Lang. Process.* 18, 2015–2027. doi: 10.1109/TASL.2010.2040793
26. Ruiz, N., & Federico, M. (2014). Assessing the impact of speech recognition errors on machine translation quality. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track* (pp. 261-274).
27. Teixeira, C. S., Moorkens, J., Turner, D., Vreeke, J., & Way, A. (2019, March). Creating a multimodal translation tool and testing machine translation integration using touch and voice. In *Informatics* (Vol. 6, No. 1, p. 13). MDPI.
28. TradDICT Learn Homepage, <https://www.traddictlearn.online/>, last accessed 2023/04/25
29. Toral, A., Wieling, M. and Way, A., 2018. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, 5, p.9.
30. Trados Studio Homepage, <https://www.trados.com/products/trados-studio/>, last accessed 2023/04/25
31. Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 697–702.
32. Wang, L., & Sun, S. (2023). Dictating translations with automatic speech recognition: Effects on translators' performance. *Frontiers in Psychology*, 14.
33. Zapata, J. (2014, November). Exploring multimodality for translator-computer interaction. In *Proceedings of the 16th International Conference on Multimodal Interaction* (pp. 339-343).
34. Zapata, J., Castilho, S., & Moorkens, J. (2017). Translation dictation vs. post-editing with cloud-based voice recognition: A pilot experiment. *Proceedings of MT Summit XVI*, 2.

National Language Technology Platform (NLTP): The Final Stage

Artūrs Vasiļevskis¹, Jānis Ziediņš², Marko Tadić³, Željka Motika⁴, Mark Fishel⁵,
Bjarni Barkarson⁶, Claudia Borg⁷, Keith Aquilina⁸ and Donatienne Spiteri⁹

¹ Tilde, Riga, Latvia

`arturs.vasilevskis@tilde.com`

² Culture Information Systems Centre, Riga, Latvia

`janis.ziedins@kis.gov.lv`

³ University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb, Croatia

`marko.tadic@ffzg.unizg.hr`

⁴ Central State Office for the Development of Digital Society, Zagreb, Croatia

`Zeljka.Motika@rdd.hr`

⁵ University of Tartu, Tartu, Estonia

`fishel@ut.ee`

⁶ Reykjavik University, Language and Voice Lab, Reykjavik, Iceland

`bjarnibar@ru.is`

⁷ University of Malta, Valletta, Malta

`claudia.borg@um.edu.mt`

⁸ Malta Information Technology Agency, Blata l-Bajda, Malta

`keith.aquilina@gov.mt`

⁹ Office of the State Advocate, Valletta, Malta

`donatienne.spiteri@stateadvocate.mt`

Abstract. The final stage and the demo of the National Language Technology Platform (NLTP) developed within the CEF action of the same name is presented in this paper. The action aims at combining the most advanced language technology tools and solutions in a new state-of-the-art, artificial-intelligence-driven, web-based national platform for language technology oriented primarily towards users from public administrations of partner states. The Platform combines into a single framework the CAT tools, the TMs usage and management, the terminology management, several different MT engines and other language technology modules.

Keywords: machine translation, CAT tools, parallel corpora.

1 Introduction

The paper presents the final stage on the CEF action National Language Technology Platform (NLTP). The general aim of the action is to combine the most advanced Language Technology (LT) tools and solutions in a new state-of-the-art, artificial-intelligence-driven, web-based national platform for LT. The action is in its final stage with the fully functional systems being deployed at the level of partner states

(Latvia, Croatia, Estonia, Iceland, and Malta). In parallel, the planned data collection has been completed and consequently the machine translation (MT) systems training is also finalised. The paper is structured as follows: in Section 2 the previous projects and related work are presented. The targeted users are described in Section 3 while the details of the development process are given in Section 4. In Section 5 we provide the information about the NLTP sustainability and possible future directions.

2 Related work

The developed solution in NLTP¹ builds on the already existing `hugo.lv` platform and the results of the *EU Council Presidency Translator* (INEA/CEF/ICT/A2018/1762093)² action, which have proven beneficial over multiple years of active use. However, these two predecessors have been substantially extended into NLTP in order to provide public administrations, SMEs and the general public with secure access to high quality MT and integration with computer aided translation (CAT) tools, e-mail and web plug-ins etc., for translation of texts, documents and web pages. At this stage the set of offered services is considered final, but the modular design of the platform allows it to be enriched with additional LT services beyond this initial set. The NLTP modules and the overall structure of the platform are presented at Figure 1. The broader context and the introduction to NLTP can be consulted in two previous papers [2, 3] where motivation and overall view has been presented.

3 Users

In its final form NLTP is adapted, localised, and sustainably deployed by the public administration bodies in partner states, while its development is supported at the same time by local research institutions as complementary partners. In the case of Iceland and Estonia, the research partners were given the role of public authorities as well. Additionally, the NLTP was customised to the specific needs of public administrations so it provides translation using our own MT systems, but it is also additionally linked to eTranslation³ services, thus enabling translations into and from the 24 official EU languages and other languages offered by eTranslation.

After the user needs were modelled following the estimated overall general needs, the additional specific requirements have been collected through a survey about LT needs and expectations in the public administration, that has been run in partner states. The example of such analysis of the survey in Croatia was presented in [1]. According to this survey e.g. 67% of users were familiar with CAT tools and 33% with MT, in 45% of institutions no LT is being used, the most useful LT in public administration are CAT tools (29%) and MT (11%).

¹ <https://nltp-info.eu>

² <https://presidencymt.eu>

³ <https://ec.europa.eu/digital-building-blocks/wikis/display/CEFDIGITAL/eTranslation>

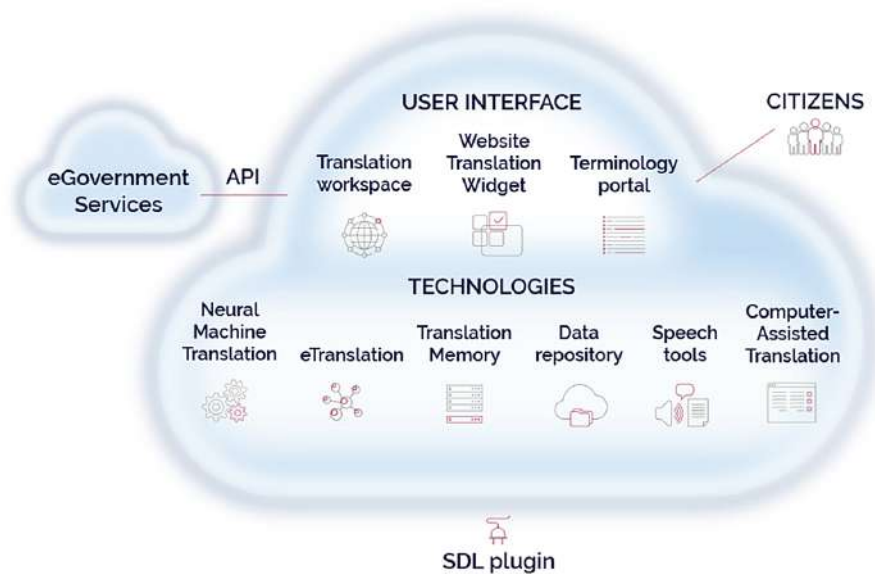


Fig. 1. The NLTP modules and overall structure.

The NLTP increases the efficiency of translation, the reuse of translation memories and the use of the existing high-quality MT technologies. Additionally, the action also integrates speech technologies for selected languages with automatic speech recognition and/or text-to-speech services. However, speech modules are not available for all languages at this moment since for some of them no such modules exist at all.

4 Development

Beside already customary services in the form of MT for pasted text and uploaded documents, the NLTP also provides additional valuable services such as a professional translation environment in the form of CAT tool integration through a simple-to-use online access in a plain browser and coupled with a number of other technological solutions, such as a translation widget, browser plugin, commercial CAT tool plugins, etc. This set of services ensures the widest possible reach to the users since these services could cover the needs of many users when they have to deal with the multilingual content. Providing public administration employees with a free, easy-to-use professional translation environment will further increase their productivity by creating a

cycle of use and reuse of translated content through translation memories (TM) accumulated in the process.

The platform also integrates services for terminology management linked with national terminology databases, as well as common European IATE⁴ terminology database.

NLTP features MT systems tailored to the specific domains of administrations following their specific language, terminology, and communication styles. Examples of domains are legal, financial, medical and other areas of public administration that feature specific use of language. Customization maximizes translation quality for the local languages of the hosting country.

Each national variant of NLTP has been adapted according to the desired visual template, the interface and the help system have been localized in the national language, while for international users the interface in English is also available.

For each national variant a technical solution was packed in a Docker⁵ installation and it has been integrated into the existing digital services at the national level.

4.1 Deployment

The platform was developed according to the common overall concept, but since the current eGovernment systems in partner states differ substantially, for each partner state a deployable variant had to be adapted to the needs of public administrations at the national level.

An example of such variant of deployment in Croatia can be seen in the Figure 2 and Figure 3. For instance, in Croatia the NLTP became an integral part of horizontal digital eGovernment services that are accessible by everyone working in the public administration at any level: national, regional and local. Also, this horizontal services are offered to anyone who has the authentication and authorisation privileges for lowest level of eGovernment services and this practically encompasses the whole public sector. There are also plans to offer these services to SMEs.

The similar deployment was conducted in other partner states, but adapted to their specific conditions and needs.

4.2 Additional datasets

Additionally, a number of domain-specific parallel datasets has been collected for five languages (Latvian, Croatian, Estonian, Icelandic, and Maltese) coupled with English. These datasets will be made available through the ELRC-SHARE⁶ repository in the Translation Memory eXchange (TMX) or similar compatible format. Since the sources of data are predominantly coming from the public domain, the data will be accessible under permissive licences.

⁴ <https://iate.europa.eu>

⁵ <https://www.docker.com>

⁶ <https://elrc-share.eu>

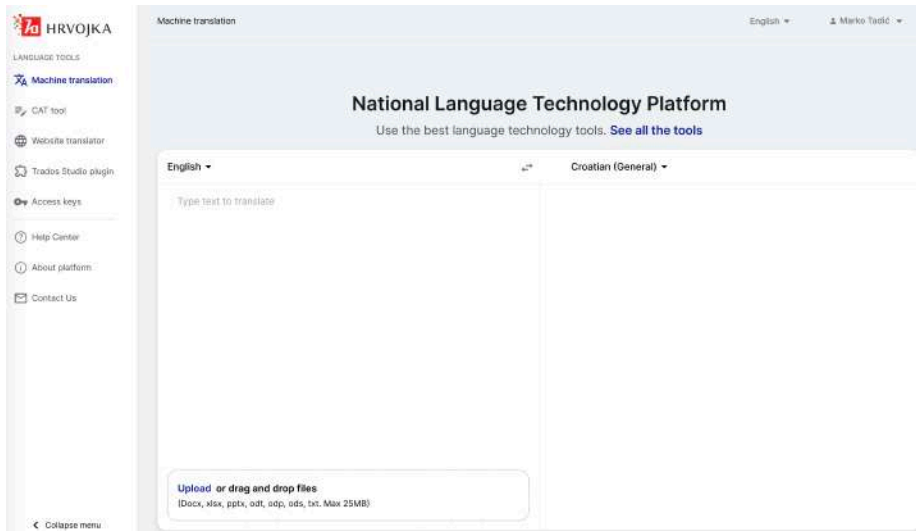


Fig. 2. Example of deployment in Croatia named Hrvojka, available at <https://hrvojka.gov.hr>. The typical MT service for translation of pasted text or uploaded document is presented with the web interface set to English for presentation clarity.

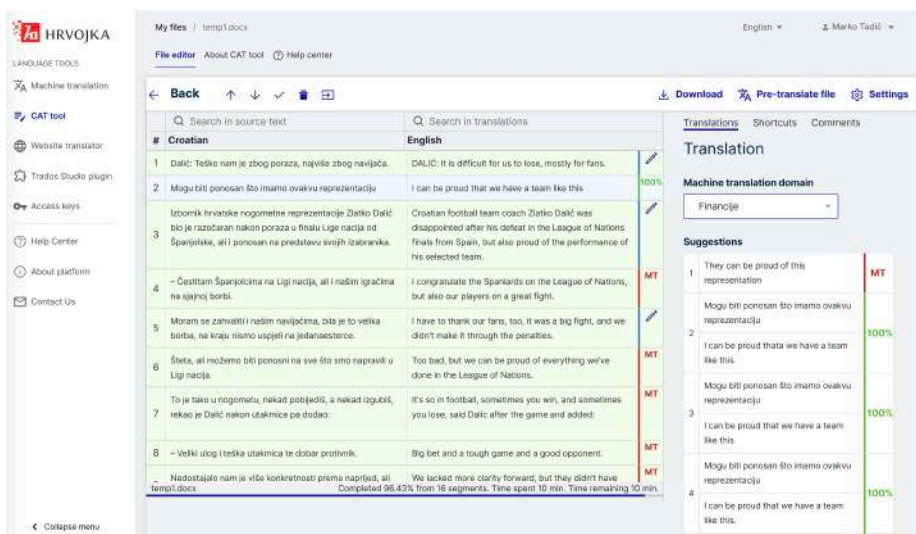


Fig. 3. Example of deployment in Croatia named Hrvojka, available at <https://hrvojka.gov.hr>. The CAT service running in a plain browser is presented with the web interface set to English for presentation clarity.

5 Sustainability and Future Directions

The public administration partner institutions are responsible for the sustainability of each national NLTP after the action ends. This has been provided by securing its inclusion into the national infrastructures for eGovernment as cloud services. This will enable multilingual access to and by public administrations, while, at the same time, the integration with public digital services offered in languages of the EU and EEA will be fostered.

For future research and development directions, similar platforms could be developed and deployed for other EU member states, and in this respect this action can be regarded as the proof-of-concept.

Also, with the introduction of new European language data sharing and processing initiatives such as Language Data Space and European Data Infrastructure Consortium being established for the language data, it is expected that services similar to NLTP would become more frequent and more readily available.

Acknowledgements

The work reported here was supported by the European Commission through the CEF Telecom Programme (Action No: 2020-EU-IA-0082, Grant Agreement No: INEA/CEF/ ICT/A2020/ 2278398, duration 2021-04-01–2023-06-30).

References

1. Motika, Ž., Didak Prekpalaj, T., Horvat Klemen, T., Koščec Perić, M.: Predstavljanje projekta *Nacionalna platforma za jezične tehnologije* (invited lecture). In: MIPRO 2022 - 45th Jubilee International Convention: CIS-AIS – Artificial Intelligence Systems, Opatija (2022), <http://www.mipro.hr/MIPRO2022.CIS-AIS/ELink.aspx>, last accessed 2023/06/19
2. Tadić, M., Farkaš, D., Filko, M., Vasiļevskis, A., Vasiļjevs, A., Ziedīņš, J., Motika, Ž., Fishel, M., Loftsson, H., Guðnason, J., Borg, C., Cortis, K., Attard, J., Spiteri, D.: National Language Technology Platform for Public Administration. In: Aldabe, I., Altuna, B., Farwell, A., Rigau, G. (eds.) Proceedings of the LREC 2022 workshop Towards Digital Language Equality (TDLE 2022), pp. 46–51. European Language Resources Agency, Marseille (2022).
3. Vasiļevskis, A., Ziedīņš, J., Tadić, M., Motika, Ž., Fishel, M., Loftsson, H., Guðnason, J., Borg, C., Cortis, K., Attard, J., Spiteri, D.: National Language Technology Platform (NLTP): overall view. In: Macken, L., Rufener, A., Van den Bogaert, J., Daems, J., Tezcan, A., Vanroy, B., Fonteyne, M., Barrault, L., Costa-Jussà, M., Kemp, E., Pilos, S., Declercq, C., Koponen, M., Forcada, M., Scarton, C., Moniz, H. (eds.) Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, pp. 343–344. European Association for Machine Translation, Ghent (2012).

Syntactic Quality Measurement in Machine Translation with Interlinguas

Alessandro Maisto¹ and Javier Oliver²

¹ University of Salerno, Via Giovanni Paolo II, 132, Fisciano (SA), Italy
amaisto@unisa.it

² Universidad Politécnica de Madrid & DAIL Software S.L., Madrid, Spain
j.oliverd@alumnos.upm.es

Abstract. The measurement of machine translation (MT) performances is an unsolved issue in NLP. This task can be done by a human but the time cost and the need for skilled workers to do it rise the necessity for automatic ways to measure the quality of a translation.

In this work, we aim to develop a new methodology for measuring the quality of MT results from a syntactic point of view. The idea takes as a theoretical framework the work of Harris about the decomposition of sentences in elementary units called kernels. Our model parses Spanish sentences and UNL (Universal Networking Language) Graphs with a rule-based methodology and divides them into units of information. Comparing those units the model measures the quality of the translation. Our results show that decomposing the sentences in minimal syntactic units could improve the evaluation performances also without a lexical/semantic analysis.

Keywords: Machine Translation Evaluation · Interlingua Translation · Syntactic Analysis

1 Introduction

Quality measurement has been an issue in the translation field for a long time. Traditionally, quality measurement has been performed by specialized workers who have spent a significant amount of time on those tasks. This issue characterised Machine Translation (MT) since its introduction. Assessing the quality of a translation is a problem that has yet to be solved in MT and NLP.

A great number of quality metrics have been introduced to evaluate MT results [21], but in many cases, they need reference translations and therefore they are not fully automatic. Nowadays, measuring sentences without their reference translation is an unresolved task.

The goal of this article is to automatically assess the quality of a translation at the syntactic level. Future studies could broaden the system to include a semantic-level measure of translation quality. We chose to measure the translation from the Spanish Language to the UNL (Universal Networking Language) interlingua, namely the U3+ implementation of the UNL language.

In the syntactic representation of sentences, each word is tagged with POS description, head dependencies and relation description. Evaluating the MT at the word level is unreliable because, in many cases, the concept could be expressed by a multiword expression or the meaning of a word could be defined only at the sentence level by its context. For example, the Spanish verb *abatir* take a different meaning when used in a psychological sense (*la película me ha abatido*, "the movie got me down") or in a concrete sense (*la grúa ha abatido el edificio*, "the crane has brought down the building").

In his work [16] Harris proposes to reduce complex sentences in kernels, i.e. elementary sentence units, to summarize scientific texts. Our idea is to set up a set of rules for sentence decomposition inspired by Harris' transformations in order to decompose the sentence into a graph of elementary structures. Those structures are larger than words and contain all the elements that allow a correct interpretation of the meaning of each linguistic element in the sentence. In addition, decomposing the original sentence and its translation with the same rules, allow us to compare their syntactic structures easily.

Our work focuses on the syntactic similarity between the input sentence and its UNL representation. Hence, testing translation quality at the syntactic level. Matusov et al. [20] provide an example of measuring the quality of a translation by dissecting the sentence.

The paper is structured as follows: in Section 2 we describe the state of the art in the field of quality measurement in MT. Section 3 illustrates what are the Universal Networking Language (UNL) and its qualities. Section 4 outlines the theoretical mark of the work. Section 5 describes the model implemented to automatically measure the quality of translation. Section 6 describes the outcomes of the experiments and in Section 6 conclusions are explained.

2 State of the art

Interlingual Machine Translation (IMT) is based on 2 steps [25]: first, the input sentence is transformed into an interlingual representation (this model is called Encoder, ENCO); second, the interlingual representation is translated to another language using a Decoder (DECO).

This ENCO-DECO model's still used nowadays for new interlingual representation. One of them is UNL (Universal Networking Language), an interlingual representation proposed by Uchida [33]. UNL is based on conceptual graphs. The nodes are Universal Words that represent concepts while the edges express semantic relations. UNL is still growing with new specifications, amplifying the range of semantics relations and the universal words' features as UNL++ [1] or U3+.

There are a few systems of machine translation based on UNL, such as Saravanan's Tamil enconverter [9], Kumar's Sanskrit enconverter [28] or Geetha's Tamil deconverter [8] and the Shi and Chen's Chinese deconverter [26].

UNL expressions can represent the meaning of a sentence in any language. Moreover, the UNL language is unambiguous. Therefore it can be decomposed

into little units of meaning which represent linguistic patterns. Cadeñosa suggested using these patterns for encyclopedic information extraction [2].

Moreover, this kind of decomposition can be useful for syntactic analysis, similarity measurement among sentences, and translation. Examples of UNL applications for translations have been presented by Mellebeck and Owczarzak [22] while Wang, Mi and Ittycheriah [34] used UNL for sentence comparison.

Nowadays, the problem of evaluating the quality of translation is the need for a canonical form to compare the results [4]. However, there are several fully Automatic Machine Translation Evaluation (AMTE) metrics. They can be classified into five categories [4]: lexical [31, 23], character [30], semantic [18, 24], syntactic [3, 13, 19, 5], and semantic-syntactic metrics [7].

Among syntactic AMTE metrics, MaxSIM [3], Helpor [13] and HWCM [19] use dependency parsing and POS tags to compute the similarity between two sentences, overlooking semantic information. Guzman proposes to add discourse structures to the quality measure metrics [11]. Guzman demonstrated the advantage of using discourse structures with kernels [10] and with neural networks [12].

3 Universal Networking Language (UNL)

UNL is a universal language based on the theory of conceptual graphs [29] which describe lexical semantics and knowledge without ambiguity.

Since it is an interlingua, UNL needs an encoder-decoder architecture. As explained by Uchida in his work [33] UNL has the following elements:

- Universal words (UW) which are the nodes of the UNL graphs. They represent a concept identified by an English word and by a set of restrictions which disambiguate it.
- Relations, which are the links between nodes. They are semantic relations which also mark the syntactic structure of the graph. There are limited types of relations such as causality, temporary, etc.
- Attributes, which specify morpho-syntactic among other features from the nodes such as number, tense, etc.

The shape of a UNL graph is depicted in figure 1, together with its textual representation.

4 Linguistic Theory

The linguistic theory of Zellig Harris introduced the concept of "transformation" as rules of correspondence between different sentences. In Harris' theory [14, 15] sentences are based on a relation among two kinds of linguistic elements: Operators and Arguments. Operators are high-level elements which determine the syntactic structure of the sentence by selecting specific arguments. For example, the operator "to eat" selects an argument which represents who eats and an

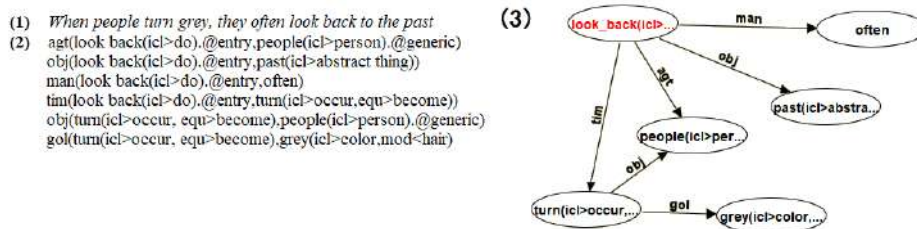


Fig. 1. Example of UNL: (1) is the original text; (2) is the UNL textual representation; (3) is the UNL Graph form

argument which indicates what is eaten. The relation among Operators and Arguments "gives a word-sequence the capacity to express fixed semantic relations among its words". It is realized into the *Kernel Sentences*, namely sentences that include only the elements required by the operator.

Harris identified a series of fundamental transformations that allow *kernels* to be derived from any sentence. In all these transformations, the partial order and the "major elements of meaning" are preserved [15, p. 290]. Basic transformations include:

- zeroing or reduction of elements deductible from the context (*John is coming but Mary is not* → *John is coming but Mary is not coming*).
- permutation of word-classes (*This the teachers deny* → *The teachers deny this*).
- single-word adjuncts (*He is very slow* → *he is slow*).
- sentence nominalization (*John's anger surprised me* → *John is angry*).
- conjoined sentences (*The concert was given inside because it was raining* → *the concert was given inside + it was raining*).

Harris identifies seven kind of kernels which can be resumed as³:

1. *NvV* in which V does not accept a direct object nor an indirect argument (*Mary sleeps, Mary is sleeping*);
2. *NvVPN* where P N is a prepositional complement which is an argument of V (*Mary goes to the park*);
3. *NvVN* (*Mary eat a pizza*);
4. *NbeN* (*John is an engineer*);
5. *NbeA* (*John is intelligent*);
6. *NbePN* (*The table is in the kitchen*);
7. *NbeD* (*The lesson was yesterday*)

Harris [17] affirmed that transformations and reduction to kernel sentences can be easily used in Information Retrieval Tasks. In fact, it is possible to compare kernel sentences to detect repeated information in texts, since reducing a

³ N = Nouns; v = auxiliary, modal, aspects verbs; V = verbs; P = prepositions, A = Adjectives; D = Adverbs

complex sentence into a kernel graph helps to find nuclear information units that, in presence of the same lexicon and syntactic structure have exactly the same meaning.

5 The proposed model

Sentence decomposition can be used to measure the similarity between sentences. Our model decomposes the input Spanish sentence and its UNL representation in kernels to see if they are similar at a syntactic level. Relations among kernels are stored in a kernels graph. In particular, our measure of similarity is calculated by comparing the number of kernels, the shape of the graph, and its relations.

5.1 Decomposition of Spanish sentences

The decomposition of the Spanish sentences in a kernels graph starts with a syntactic analysis of the sentence. We used Stanford's CoreNLP Parser [6] to perform the syntactic analysis. Apparently, the use of a parser subordinates the results of our model to the precision of the parser. We chose a selected group of sentences for which the syntactic analysis did not present errors to exclude the influence of parsing errors in the evaluation of our metric.

The sentence decomposition is subdivided into four main processes that work in sequence:

1. Unification of group of words
2. Identification of subordinate sentences
3. Identification of kernel structures
4. Generation of kernels' graph

Unification of group of words In this step, we select terms that are syntactically dependent on the same head. We unify in the same group the CoNLL rows of nouns, verbs or prepositions with the elements that depend on them. With this operation, we will be able to use those groups as arguments in the following steps.

Starting from the identification of the heads of those groups, we employed syntactic relations to detect connected elements of the same group. Table 1 shows the syntactic relationship that characterizes the linguistic pieces that will be associated with each head. The heads are identified by POS tags while connected words by dependence descriptors (and, rarely, NOUN and SYM POS tags). Adverbs are always attached to the group. The words that belong to the same group are then merged together. The group inherits morpho-syntactic and semantic information from the HEAD.

For example, the Spanish sentence *la lectura del libro fue rapida* ('the reading of the book was fast') is represented in the CoNLL structure illustrated in table 2.

Table 1. Rules of group-words unification: for a description of the Universal Dependency Relation and POS tags consult <https://universaldependencies.org/format.html>

Head	det	amod	case	nummod	aux	cop	NOUN	SYM	fixed
NOUN	x	x	x	x					
NUM	x	x	x	x					
PRON	x	x	x	x					
PROPN	x	x	x	x					
VERB					x	x			
ADJ					x	x			
All							x	x	x

Table 2. CoNLL of the sentence *la lectura del libro fue rapida* before the Unification of group of words

ID	Token	POS	HEAD	DEP
1	la	DET	2	det
2	lectura	NOUN	6	nsubj
3	del	ADP	4	case
4	libro	NOUN	2	nmod
5	fue	AUX	6	aux
6	rapida	ADJ	0	ROOT

Table 3. The CoNLL of *la lectura del libro fue rapida* after the Unification of group of words

ID	Token	POS	HEAD	DEP
2	la lectura	NP	6	nsubj
4	del libro	PP	2	nmod
6	fue rapida	VP	0	ROOT

The system takes as HEADs the Nouns *lectura* 'reading', and *libro* 'book', and the Adjective *rapida* 'fast'. The noun *lectura* will be unified to its determiner *la* 'the' because they are related by a *det* arc. The final result is a Noun Phrase (NP). The resulting structure is shown in table 4.

Identification of subordinate sentences The second step regards the recognition of subordinates from the main clause. Our system detects subordinates, splits the original sentence into different periods, and recognizes the type of subordination involved. We select Subjective, Objective, Prepositional/Adverbial and Relative clauses.

Taking the verb as a marker, the system captures all the elements involved in the subordinate by searching connected linguistic elements. Then, syntactic relations are used to classify the sentence. We use *csbj* relations for subjective clauses, *ccomp* for objective clauses and *advcl*, *acl*, *xcomp* and *conj* for all the other subordinate clauses⁴.

The input for the sentence *el insecto que aplasté era una cucaracha* 'the insect I squashed was a cockroach' is shown in table 4.

The system loop searches for a subordinate marker in the DEP column of the sentence until it finds the *acl* in row 4. All the elements pointing to the element with ID 4 are then stored in a list. Subsequently, the same operation is

⁴ <https://universaldependencies.org/u/dep/index.html>

Table 4. Structure of Sentence *el insecto que aplasté era una cucaracha*

ID	Token	POS	HEAD	DEP
2	el insecto	NP	7	nsubj
3	que	PRON	4	nsubj
4	aplasté	VERB	2	acl
5	era	AUX	7	cop
7	una cucaracha	NP	0	ROOT

repeated for each new element of that list. When there are no new elements in the variable, the algorithm stops and generates a sentence for each subordinate. In the example above, the elements are the relative sentence *que aplasté* and the main sentence *el insecto era una cucaracha*.

Identification of kernel structures To detect what is part of a kernel sentence (an argument of a specific verb), we need a list of verb structures that specify the number and the kind of arguments they select.

For our experiment, we manually extract a list of 2934 Spanish verbs each of which was annotated with its essential syntactic structure. The structures were the following:

- 0V (12 intransitive verbs with no complements or subject as "llover", *to rain*)
- NV (373 intransitive verbs with subject as "dormir", *to sleep*)
- NVN (1866 transitive verbs with the direct object as "comer", *to eat*)
- NVPN (59 intransitive verbs with one prepositional complement as "mentir", *to lie*)
- NVNPN (624 transitive verbs with a direct object and a prepositional complement as "decir", *to say*)

Since the Spanish version of the Stanford Parser does not perform Lemmatization⁵, we automatically inflect the terms in the list to associate the structures with all the inflected forms of each verb. We use the software NooJ [27], which provides basic resources for the Spanish language, in order to generate a complete inflected dictionary of verbs.

For each sentence, the algorithm searches for the verb and extracts its argument structure from the dictionary. This structure indicates what complement is not an argument, following the rules shown in Table 5.

If we take as an example the sentence "Antonio durmió todo el día en la cama" (*Antonio slept all day in the bed*), the algorithm detects three kernel sentences: a) *Antonio durmió* (*Antonio slept*), because the verb *to sleep* have an NV structure; b) *(el dormir fue) todo el día* (*(the slept was) all day*) and c) (el

⁵ <https://stanfordnlp.github.io/CoreNLP/human-languages.html>

Table 5. Rules of Argument Selection: for a description of the Universal Dependency Relation Tags consult <https://universaldependencies.org/u/dep/index.html>

Structure	obj	nmod	obl	iobj	advmod	appos
OV/NV	x	x	x	x	x	x
NVN		x	x	x	x	x
NVPN	x				x	x
NVNP			x		x	x

dormir fue) en la cama ((*the slept was*) in the bed), because they represent two *obl* complement.

Generation of Kernel’s Graph All the kernels generated are organized as a graph. Each node represents a kernel which is connected to another kernel by a relation in an expression of the kind *RELATION(A,B)*. The list of relations is the following:

- SUBJ arcs, which includes subjective sentences (A: "me entristece" (*it saddens me*); B: "que te vayas" (*that you are leaving*); SUBJ(B,A));
- OBJ arcs, which includes objective sentences (A: "Antonio dijo" (*Antonio said*); B: "que le gustaban los coches" (*that he like cars*); OBJ(B,A));
- COMP arcs, which includes prepositional sentences (A: "Antonio iba" (*Antonio was going*); "a coger el coche" (*to get the car*) COMP(B,A));
- AGG arcs, which includes circumstantial complements (A: "me comeré una pizza" (*I will eat a pizza*); B: "en mi restaurante favorito" (*in my favourite restaurant*); AGG(B,A));
- ARC arcs, which includes adjectival, prepositional or adverbial phrases dependent from a non-verbal phrase (A: "la lectura fue rápida" (*the reading was fast*); B "(la lectura fue) del libro" (*(the reading was of the book)*); ARC(B,A));
- REL arcs, which includes relative sentences (A: "el restaurante" (*the restaurant*); B: "que más me gusta" (*that I like the most*); REL(B,A));
- MARK arcs, which includes modal, causal, final or other subordinate sentences (A: "iba hacia el coche" (*I was walking towards the car*); B: "cuando oí un disparo" (*when I heard the shot*); MARK(A,B));
- CONJ arcs, which includes coordinated sentences or phrases (A: "Antonio dormía" (*Antonio has slept*); B: "mientras su mujer sonaba" (*while his wife played*); CONJ(A,B));

The final graph includes kernels (the linguistic units) and relations. In the next sections, we will illustrate how the comparison with the UNL kernel’s graphs is performed and how we score the translation precision score.

5.2 Decomposition of UNL sentences

The decomposition of the UNL sentences is done by a shallow parsing technique, guided by the analysis of the sentence with pattern detection. The shallow pars-

ing technique is based on a knowledge base of patterns which is processed by a pattern searcher. The patterns of the knowledge base follow the next structure:

- A first part where we declare the content of the pattern's nodes.
- A second part where we declare the relationships between nodes and the nodes they relate.

As an input, the method receives a UNL graph in text form as in the figure 1. The aim is to decompose the graph in patterns to observe its internal structure. This structure represents the syntactic structure of the UNL graph because it shows the relations between the Universal Words. In figure 1, for example, the relations of the predicate *look back* are expressed by an *agt* (agent) relation (*people*), a *obj* (object) relation (*past*) and other modifiers (*man* (manner) - *often*, *tim* (time) - *turn* which includes another sentence).

Decomposition of the UNL Graph In the decomposition steps, the system searches for the verbs' complements to detect the syntactic structure of the sentence. For example, the pattern that searches for the structure of a type *be* verb relates the nodes 2 and 3 which represent its subject and object 2.

```
1:.*?>be.*?).*?
   aoj(:.*)?(1,2)
   obj(:.*)?(1,3)
```

This pattern would represent the graph of figure 2.

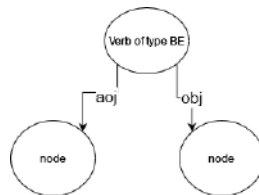


Fig. 2. Graph pattern

Redundant Information Pruning At the end of the detection phase, the information enclosed in the captured patterns can overlap. For example, we might find a pattern which includes a subject and an object (aoj-obj) and another pattern which includes the same subject and object with an object modifier (aoj-obj-mod). In this case, we include in the list of patterns only the one which contains more information.

Special patterns As mentioned before, we aim to capture the biggest patterns in the expression. Nevertheless, a smaller pattern can be part of a bigger one. In this case, we planned a specific step in which we split the bigger patterns into more units.

Generation of the Graph Once we have all the occurring patterns in the input graph we relate them to create the rest of the output. The relations between the patterns are the same used for the decomposition of the Spanish sentences. We use the same output model explained in the section 5.1. We give an incremental letter to each relation found as an identifier (ID). In order to relate those kernels we look to the first or last node of every kernel. Kernels with the same first or last node are related. Other relations such as MARK are done by searching specific relations inside the UNL structure.

We find all possible patterns inside the graph, as shown in the example in Appendix 8. The steps employed in the process are the following:

1. the system searches for the patterns in the graph.
2. After that the system prunes the redundant information from the list of kernels.
3. The list of matches is completed with the special patterns, created to solve conflicts that arise with the growing up of the knowledge base.
4. Remaining kernels are associated with each other.
5. Finally, the system orders the list of kernels by their ID.

5.3 Measuring the likeness between graphs

After breaking a text into kernels, we must compare those kernels to determine their similarity.

To compare the sentences we take both graphs (Spanish and UNL) and compare them using a metric created for that purpose. The metrics take into account the number of kernels detected, the relations in which they are involved and the name of the arcs of that relations.

The metric used is based on the mean of three indices:

1. NK : the number of total standard kernels
2. S : the shape of the graph
3. C : coherence of the relation tags

The NK index is calculated as the ratio between the number of total standard kernels contained in the two sentences.

The S index takes the shape of the graph into account as an index of correspondence between the two sentence graphs.

$$S = \frac{N - D}{N} \quad (1)$$

S is a 0-to-1 index in which the Relation Difference D measures the distance between the two graphs. N represents the number of detected kernels of the

bigger graph. With $k1_0...k1_n$ as the number of kernels that appear zero or more times in the first graph and $k2_0...k2_n$ as the number of kernels that appear zero or more times in the second graph D is calculated as:

$$D = \sum_{i=0}^n \frac{|k1_i - k2_i|}{2} \quad (2)$$

The difference between the two graphs is calculated as the number of relations in which each kernel participates. In this way, we don't need to know the nature of each kernel, but we may compare their relationships through the graph. Because each difference affects the measure twice, the final value is divided by two. The index is calculated as the correct percentage of the kernel's graph.

The C index focuses on the edge typology that connects the kernels. We distinguish four types of relations:

- Type 1: relationship of the kind *ARC*, *MARK*, and *REL*. Noun and verb modifiers of any kind are included in these relationships. Relative sentences are also included because they can be translated as a nominal modifier in some cases.
- Type 2: *OBJ*, *COMP*, and *AGG* relations. This list includes both simple and complex verb complements. We also included objective sentences because the syntactic structure of the verb might be transitive in one language and intransitive in another (for example, the verb *to call*, "llamar", in *Mark called John by telephone* correspond to "Mark llamó a John por teléfono" in Spanish).
- Type 3: *CONJ*.
- Type 4: *SUBJ*.

The formula of C is the same as that used for the formula of S , but N denotes the maximum number of edges on the graph, and the score for D is determined differently:

$$D = \sum_{i=0}^n \frac{(|k1_{t1} - k2_{t1}|) * 0.3 + (|k1_{t2} - k2_{t2}|) * 0.5 + (|k1_{t3} - k2_{t3}|) + (|k1_{t4} - k2_{t4}|)}{2} \quad (3)$$

We assigned a confidence value to each of the four types of relations identified: type 1 (t1) has the lowest weight (0.3) because these types of relations are relatively common and their recognition is frequently erroneous. The confidence weight for type 2 (t2) relationships is 0.5.

Because subjective sentences and conjunctions substantially characterize the sentences, type 3 (t3) and type 4 (t4) have a weight of one. Furthermore, the decomposition process does not make mistakes in recognizing those structures.

To clarify the stages involved in calculating the final value of similarity we will take as an example the sentence "El hermano de Antonio llamó a la policía mientras se escapaba" (*Antonio's brother called the police while he was running away*). The Spanish sentence will be divided into the following list of kernels:

1. A: el hermano llamó a la policia
2. B: el hermano es de Antonio
3. C: llamó mientras se escapaba
4. R1: ARC(B,A)
5. R2: COMP(C,A)

Assuming that the translation in UNL was the same as the one in English that follows:

1. A: the brother called the police
2. B: the brother was Antonio's
3. C: while he was running
4. D: the running is away
5. R1: ARC(B,A)
6. R2: COMP(C,A)
7. R3: ARC(D,C)

the ratio between the number of total standard kernels is equal to $3/4 = 0.75$. To calculate the relation difference D for the index S we must look at the graph in table 6. The value D is equal to 0.5 and the difference between the shape of the two graphs is equal to $(4 - 0.5)/4 = 0.875$.

Table 6. Number of labels per number of relations

Relations	Spanish	UNL	Difference
0	0	0	0
1	2	2	0
2	1	2	1

Table 7. Type of relations per number of relations

Relations	Spanish	UNL	Difference
t1	1	2	1
t2	1	1	0
t3	0	0	0
t4	0	0	0

For the index C we calculated the number of relationships that belongs to the four families (table 7). The score of D in this case is $(4 - 0.15)/4 = 0.98$.

The equivalence score between the two sentences is the mean of the three values: 0.867.

6 Experimentation

In order to test the proposed model we ran an experiment in which we calculate our metric across 100 sentences collected from the ELRC-1082-CNIO v1 Corpus [32], a corpus of Spanish scientific paper translated into English. The corpus was converted to UNL language and processed as explained in previous sections.

The sentences used in this experiment have been extracted from a corpus of Spanish sentences labelled with their validated and correct UNL expression.

To assess the system’s ability to detect incorrect translations, we assessed the model’s precision over 200 sentence pairs, half of which were associated with the correspondent translation and the other half with a random text. To determine when a sentence is correctly translated, we must first create a threshold above which we can consider the translation to be correct. We examined four alternative thresholds in this experiment: 0.8, 0.75, 0.7, and 0.6.

The precision of the methodology is shown in table 8. We took as correct every correctly translated pair with a score higher than the threshold and every random pair whose score is below.

Table 8. Precision of the Methodology over truly translated sentence pairs and random pairs

threshold	True	Random	total
0.8	0.57	0.80	0.691
0.75	0.75	0.69	0.724
0.7	0.85	0.59	0.724
0.6	0.95	0.32	0.640

Table 8 shows that a threshold of 0.7 or 0.75 yields the best overall value. Despite the fact that we did not include the lexicon in our analysis, the model gets outstanding precision results. Especially with properly translated sentences.

The study of the results reveals that in many situations, a mistaken judgment on correct translation must be attributed to an incorrect parsing result. For example, the sentence "*Utilizar el siguiente formulario para contactar con Elena*" is translated correctly to UNL but receives a score of 0.67. In fact, the analysis of the Spanish sentence evidences the presence of 3 kernels, one of which is included in two edges and two included in only one edge. In UNL, there is only one kernel.

Analyzing the incorrect pairs, the errors’ motivations could be chance, but they could also be parsing or decomposition faults. The sentence "Esto podría suponer un coste adicional y un retraso en la tramitación." (*This could reflect an additional cost and delay in the entire process*) (4 kernels), for example, achieves a score of 0.88 with the sentence "nuestros donantes tienen la oportunidad de disfrutar mientras realizan una contribución a buen fin" (*our donors have the opportunity to enjoy their contribution for a good cause*) (4 kernels). In table 9 we report the numerical analysis of the two sentences.

The two sentences generate the following kernel graphs:

- | | |
|------------------------------------|---------------------------------------|
| – k1: Esto podría suponer un coste | – k1: our donors have the opportunity |
| – k2: adicional | – k2: to enjoy their contribution |
| – k3: y un retraso | – k3: for a cause |
| – k4: en la tramitación | – k4: good |
| – r1 : ARC(k2,k1) | – r1 : OBJ(k2,k1) |
| – r2 : CONJ(k3,k1) | – r2 : ARC(k3,k2) |
| – r3 : MARK(k4,k3) | – r3 : ARC(k4,k3) |

Table 9. Comparing two not-related sentences

Number of relations	Sentence 1	Sentence 2	Difference
0	0	0	0
1	2	2	0
2	2	2	0
t1	2	2	0
t2	0	1	1
t3	1	0	1
t4	0	0	0

Since the sentence graph has the same number of kernels and the same shape, the score of NK and S are equal to 1.

Because the phrases differ in one element, the index C is equal to 0.66. Despite the fact that the two statements are made up of distinct words and have different meanings, they have nearly the same syntactic structure. We will strive to tackle this problem in the future by including a lexical similarity value.

7 Conclusions and future work

In this paper, we presented a syntactic strategy for automatically comparing translations without the need for a corpus of canonical forms or human labour. The results of the experiments reveal that the metrics can measure the likelihood of sentences in different languages at the syntactic level with a precision of 0.74.

Measuring merely syntactic similarities between two sentences can not avoid casual errors, because two sentences with the same structure can have entirely different meanings depending on the lexicon they contain.

For this reason, we shall expand the proposed metrics to include semantics. Kernel decomposition will also help with lexical/semantic analysis. In fact, kernels are language constructions that give the bare minimum of context and fixed relations among words. Using a lexical approach, we can compare kernels more easily than complex sentences because the syntactic elements in kernels must be the same number and type.

Furthermore, in order to enhance the present measure, we must address the reliance on the parsing mechanism. To avoid future failures, we will use a new method that does not rely on external resources. In fact, as we showed in Section 6, some of the problems are the result of parsing issues.

We must also compare our model to other commonly used state-of-the-art models and, if possible, improve our results by combining our technique with additional metrics such as Guzman’s tree structure comparison [11].

References

1. Boitet, C., Boguslavskij, I.M., Cardenosa, J.: An evaluation of unl usability for high quality multilingualization and projections for a future unl++ language. In:

- Gelbukh, A. (ed.) *Computational Linguistics and Intelligent Text Processing*. pp. 361–373. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
2. Cardenosa, J., Villa, M.Á.d.l., Gallardo, C.: Linguistic patterns for encyclopaedic information extraction. In: *International Conference on Flexible Query Answering Systems*. pp. 661–670. Springer (2013)
 3. Chan, Y.S., Ng, H.T.: Maxsim: A maximum similarity metric for machine translation evaluation. In: *Proceedings of ACL-08: HLT*. pp. 55–62 (2008)
 4. Chauhan, S., Daniel, P.: A comprehensive survey on various fully automatic machine translation evaluation metrics. *Neural Processing Letters* pp. 1–55 (2022)
 5. Chauhan, S., Daniel, P., Mishra, A., Kumar, A.: Adableu: A modified bleu score for morphologically rich languages. *IETE Journal of Research* pp. 1–12 (2021)
 6. Chen, D., Manning, C.D.: A fast and accurate dependency parser using neural networks. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pp. 740–750 (2014)
 7. Comelles, E., Atserias, J.: Verta: a linguistic approach to automatic machine translation evaluation. *Language Resources and Evaluation* **53**, 57–86 (2019)
 8. Dhanabalan, T., Geetha, T.: Unl deconverter for tamil. In: *International Conference on the Convergences of Knowledge, Culture, Language and Information Technologies* (2003)
 9. Dhanabalan, T., Saravanan, K., Geetha, T.: Tamil to unl enconverter. In: *Proc. International Conference on Universal Knowledge and Language*, Goa, India (2002)
 10. Guzmán, F., Joty, S., Márquez, L., Moschitti, A., Nakov, P., Nicosia, M.: Learning to differentiate better from worse translations. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 214–220 (2014)
 11. Guzmán, F., Joty, S., Márquez, L., Nakov, P.: Using discourse structure improves machine translation evaluation. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 687–698 (2014)
 12. Guzmán, F., Joty, S., Márquez, L., Nakov, P.: Machine translation evaluation with neural networks. *Computer Speech & Language* **45**, 180–200 (2017)
 13. Han, A.L.F., Wong, D.F., Chao, L.S., He, L., Lu, Y., Xing, J., Zeng, X.: Language-independent model for machine translation evaluation with reinforced factors. In: *Proceedings of Machine Translation Summit XIV: Posters* (2013)
 14. Harris, Z.: *Language and information*. Columbia University Press (1988)
 15. Harris, Z.: *Theory of language and information: a mathematical approach* (1991)
 16. Harris, Z.S.: Discourse analysis. *Language* **28**(1), 1–30 (1952)
 17. Harris, Z.S.: Linguistic transformations for information retrieval. *Papers in structural and transformational linguistics* pp. 458–471 (1970)
 18. Li, P., Chen, C., Zheng, W., Deng, Y., Ye, F., Zheng, Z.: Std: An automatic evaluation metric for machine translation based on word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **27**(10), 1497–1506 (2019)
 19. Liu, D., Gildea, D.: Syntactic features for evaluation of machine translation. In: *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. pp. 25–32 (2005)
 20. Matusov, E., Leusch, G., Bender, O., Ney, H.: *Evaluating machine translation output with automatic sentence segmentation* (2005)
 21. Maučec, M.S., Donaj, G.: Machine translation and the evaluation of its quality. *Recent Trends in Computational Intelligence* p. 143 (2019)

22. Mellebeek, B., Owczarzak, K., Van Genabith, J., Way, A.: Multi-engine machine translation by recursive sentence decomposition. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers. pp. 110–118 (2006)
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
24. Rei, R., Stewart, C., Farinha, A.C., Lavie, A.: Comet: A neural framework for mt evaluation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2685–2702 (2020)
25. Richens, R.H.: Interlingual machine translation. *The Computer Journal* **1**(3), 144–147 (1958)
26. Shi, X., Chen, Y.: A unl deconverter for chinese. *UNL Book* (2005)
27. Silberztein, M.: Nooj’s dictionaries. *Proceedings of LTC* **5**, 291–295 (2005)
28. Sitender, Bawa, S.: Sansunl: a sanskrit to unl enconverter system. *IETE Journal of Research* **67**(1), 117–128 (2021)
29. Sowa, J.F.: Conceptual graphs. *Foundations of Artificial Intelligence* **3**, 213–237 (2008)
30. Stanchev, P., Wang, W., Ney, H.: Eed: Extended edit distance measure for machine translation. In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). pp. 514–520 (2019)
31. Su, K.Y., Wu, M.W., Chang, J.S.: A new quantitative quality measure for machine translation systems. In: *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics* (1992)
32. Tiedemann, J.: Parallel data, tools and interfaces in opus. In: *Lrec. vol. 2012*, pp. 2214–2218. Citeseer (2012)
33. Uchida, H.: The UNL specification. <http://www.unl.org/> (2003), [Online; accessed 23-April-2022]
34. Wang, Z., Mi, H., Ittycheriah, A.: Sentence similarity learning by lexical decomposition and composition. *arXiv preprint arXiv:1602.07019* (2016)

8 Appendix I: UNL Decomposition example

The sentence "Estamos interesados en seguir con la campaña otro mes" is represented in UNL as:

```

aoj(be(icl>be,aoj>thing,obj>thing).@entry.@present, we(iof>person))
obj(be(icl>be,aoj>thing,obj>thing).@entry.@present,
interested(icl>adj,com>interest,obj>thing))
obj(interested(icl>adj,com>interest,obj>thing),
continue(icl>act>do, equ>go_on, agt>thing, obj>thing, cob>thing))
obj(continue(icl>act>do, equ>go_on, agt>thing, obj>thing, cob>thing),
campaign(icl>operation>thing).@def)
tim(continue(icl>act>do, equ>go_on, agt>thing, obj>thing, cob>thing),
month(icl>period>time>thing, pof>year))
mod(month(icl>period>time>thing, pof>year), another(icl>adj))

```

Step 1: we detect all the possible patterns of the UNL graph. Since there are too many we show three of them as an example:


```

(1)
aoj(be(icl>be,aoj>thing,obj>thing).@entry.@present, we(iof>person))
obj(be(icl>be,aoj>thing,obj>thing).@entry.@present,
interested(icl>adj,com>interest,obj>thing))
(2)
obj(be(icl>be,aoj>thing,obj>thing).@entry.@present,
interested(icl>adj,com>interest,obj>thing))
obj(interested(icl>adj,com>interest,obj>thing),
continue(icl>act>do, equ>go_on, agt>thing, obj>thing, cob>thing))
(3)
obj(be(icl>be,aoj>thing,obj>thing).@entry.@present,
interested(icl>adj,com>interest,obj>thing))
obj(interested(icl>adj,com>interest,obj>thing),
continue(icl>act>do, equ>go_on, agt>thing, obj>thing, cob>thing))
obj(continue(icl>act>do, equ>go_on, agt>thing, obj>thing, cob>thing),
campaign(icl>operation>thing).@def)

```

Steps 2 and 3: we delete pattern 2 which is redundant (pruning). Then, we detect a special pattern in object 3 and we split it into two patterns.

```

(1)
aoj(be(icl>be,aoj>thing,obj>thing).@entry.@present, we(iof>person))
obj(be(icl>be,aoj>thing,obj>thing).@entry.@present,
interested(icl>adj,com>interest,obj>thing))
(3)
obj(be(icl>be,aoj>thing,obj>thing).@entry.@present,
interested(icl>adj,com>interest,obj>thing))
obj(interested(icl>adj,com>interest,obj>thing),
continue(icl>act>do, equ>go_on, agt>thing, obj>thing, cob>thing))
(4)
obj(interested(icl>adj,com>interest,obj>thing),
continue(icl>act>do, equ>go_on, agt>thing, obj>thing, cob>thing))
obj(continue(icl>act>do, equ>go_on, agt>thing, obj>thing, cob>thing),
campaign(icl>operation>thing).@def)

```

Steps 4 and 5: we search for relations between patterns. then, we write them as kernels in order to produce a labelled output:

```

"A": "aoj(be(icl>be,aoj>thing,obj>thing).@entry.@present,we(iof>person))
obj(be(icl>be,aoj>thing,obj>thing).@entry.@present,
interested(icl>adj,com>interest,obj>thing))",
"B": "obj(be(icl>be,aoj>thing,obj>thing).@entry.@present,
interested(icl>adj,com>interest,obj>thing))
obj(interested(icl>adj,com>interest,obj>thing),
continue(icl>act>do, equ>go_on, agt>thing, obj>thing, cob>thing))\n",
"C": "obj(interested(icl>adj,com>interest,obj>thing),
continue(icl>act>do, equ>go_on, agt>thing, obj>thing, cob>thing))

```

```
obj(continue(icl>act>do, equ>go_on, agt>thing, obj>thing, cob>thing),
campaign(icl>operation>thing).@def)\n",
"D": "obj(interested(icl>adj, com>interest, obj>thing),
continue(icl>act>do, equ>go_on, agt>thing, obj>thing, cob>thing))
tim(continue(icl>act>do, equ>go_on, agt>thing, obj>thing, cob>thing),
month(icl>period>time>thing, pof>year))
mod(month(icl>period>time>thing, pof>year), another(icl>adj))\n",
"E": "ARC(A,B)",
"F": "ARC(B,C)",
"G": "ARC(C,D)"
```

Analysis And Evaluation Of ChatGPT-Induced HCI Shifts In The Digitalised Translation Process

Pilar Sánchez-Gijón^[0000-0001-5919-4629] and Leire Palenzuela-Badiola¹

¹ Universitat Autònoma de Barcelona, 08192 Bellaterra, Spain

Abstract. The appearance of LLM such as ChatGPT challenges the modelling of the translation digital process in phases, tasks, and tools for didactic purposes. This article examines ChatGPT's performance in carrying out each task of the translation process using the Tradumàtica model. The findings allow us to confirm the changes in HCI that have already begun to take place, as well as how to begin incorporating them in a review of the digitalized translation process.

Keywords: Translation model, ChatGPT, Chain-of-Thought-Prompt, HCI.

1 Introduction

The appearance of the Large Language Model (LLM) ChatGPT, which is web-based and looks to be capable of performing a wide range of language-related tasks, compels us to reconsider the digitised translation process, the tasks involved, and how they are carried out. The ability to interact with GPT3 as a chatbot, just as any other intelligent home assistant, appears to have new outcomes, and so does ChatGPT 4. The purpose of this paper is to investigate and evaluate the impact of LLMs such as ChatGPT on Human-Computer Interaction (HCI) in translation environments. To that end, we conduct a systematic study based on Tradumàtica's digitised translation process model [1] to determine whether this or future LLMs will be able to: 1) correctly process any translation-related request, 2) provide a response tailored to the nature of the request, and 3) provide an adequate and satisfactory response.

2 Background

Since the early days of translation studies, both academically and professionally, an effort has been made to structure and categorise all the field's concepts. From Holmes' classification of the field into different branches of study [2] to the formalisation of translation competence [3] and [4], all classification proposals share the same goal: to delimit the field, establish possible links between the theoretical and academic approaches and professional practise.

Technology also has an impact on translation. However, the relationship between technologies and translation has primarily taken shape in these two ways: translation

hierarchies based on technological and human participation, as well as classifications of various technological resources based on their contribution to the translation process.

2.1 Classifications Of Technological Resources Regarding The Translation Process

The classification of translation tools has been a popular topic in the literature. Alcina [5] defines the various types of translation software that are currently available (machine translation, computer-assisted translation, as well as terminology managers and other tools and resources). Other authors have delved into how translators use and accept technologies [6], or into the training of translators with technologies [7] [8] based on classifications like the one suggested by Alcina. An observational exercise can also establish a direct relationship between the emergence of some of these technologies and an increase in productivity in the translation industry [9].

2.2 Classifications Of Translation Technologies And Human Interaction

Nirenburg [10] distinguishes between two Translation Workstation Systems (TWS) configurations: machine-aided human translation (MAHT) and human-aided machine translation (HAMT). TWS-MAHT was based on the following premise: "Text translated by machine can be displayed in the target text window, and the posteditor can work on it using the same tools as a translator would" [10]. TWS-HAMT, on the other hand, is based on a different principle, as defined by Nirenburg:

"[W]hile a computer tries to translate a document, a human translator monitors its progress and gives the system guidance when called upon to do so. The primary role of the human is a) to help the system make processing choices for which it lacks sufficient knowledge and, b) to update the lexicons to cover unexpected vocabulary." [10]

Wilss [11], as Kay [12] did earlier, adds a new category, referring to machine-aided human translation (MAHT), human-aided machine translation (HAMT), and fully automatic machine translation (FAMT). If we add human translation (without any kind of automated assistance) to this classification, we get the four possible scenarios that have allowed us to classify any interaction between human and machine translation.

These authors indirectly established the standard framework of HCI in translation by differentiating scenarios in which the translator's action varies in relation to the level of presence of technology. Following that, this four-level scale was gradually supplemented with hybrid forms. On the one hand, current translation projects require different degrees of human intervention even though they use the same technological resources (see [13] for the distinction between light and full post-editing). In other cases, the technological advancement of TWS has resulted in some of the differences between scenarios becoming blurred (for example, the assembly function of Translation Edit and Management Systems - TEMS, which blurs the distinction between MAHT and HAMT).

2.3 Large Language Models And ChatGPT

A Language Model (LM) is a deep learning system thought to perform natural language processing tasks regarding language [14]. Large Language Models (LLM) are a particular type of LM consisting of a neural network based on thousands of parameters from huge amounts of unlabelled text [15]. LLM are not just language-oriented, they can assume a wider range of tasks.

Several studies have been published on how such systems can change the current game rules in translation. Peng *et al.* [16] demonstrate that ChatGPT's performance is comparable to the best translation systems on the market for the most common language combinations, but that it has limitations when translating between less common language combinations. This limitation can also be found in most commercial systems. In fact, Hedy [17] achieves comparable translation output quality results and concludes that GPT translation performance is comparable to current neural machine translation systems.

The performance of LLM in translating entire documents has also been assessed [18]. While the system performs well at the sentence level, the authors conclude that it still makes discourse errors at the paragraph and document levels. Other studies have concentrated on tasks other than translation itself that are involved in the translation process. Lu *et al.* [19] focus on the possibility of using ChatGPT as a translation evaluation system, even implementing the Multidimensional Quality Metrics (MQM) metric. This study by Lu *et al.* is based on the principle of "Chain-of-Thought Prompting" [20], which highlights the reasoning ability of LLMs to perform more complex natural language processing tasks.

To the best of our knowledge, all published studies have primarily focused on assessing the quality of GPT response to tasks such as translation or translation assessment. However, the provision of translation services as defined in ISO 17100 [21] includes additional tasks required to complete a translation assignment. To conduct a systematic and comprehensive analysis of the effect of GPT in each of the translation tasks, we will follow the Tradumàtica's digitised translation process model [1] and investigate the feasibility and degree of success of "Chain-of-Thought Prompting" in each of the tasks described.

2.4 The Tradumàtica Model

The Tradumàtica digitized translation process model is a didactic model for describing and managing any professional translation process. The model's basic structure (fig. 1) depicts the digitized process as a series of phases.

These phases correspond to the process description in the ISO 17100 standard. Following this standard, the model describes all the tasks that may take place in each of the phases described. Depending on the nature of the translation brief, the resources involved, and the product to be translated, each of these tasks can be activated or bypassed.

This didactic model's goal is to assist translation students in becoming acquainted with translation technologies as well as general technologies used in translation

environments. To do so, the model provides the details required to understand what happens in each phase, what tasks must be taken to successfully progress to the next phase, and what human and technological intervention is involved in each of the possible scenarios.

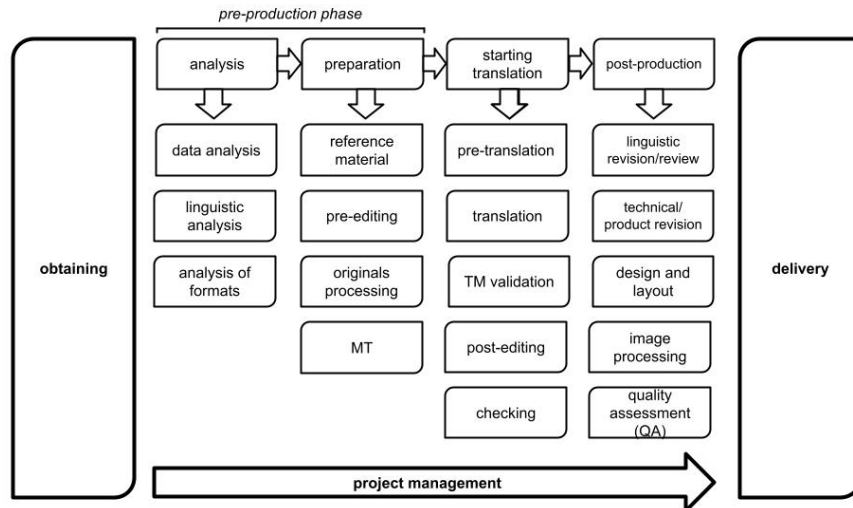


Fig. 1. Phases of the Tradumàtica model (translated into English by the authors of this paper).

In each case, a description of the need to carry out a specific task is provided, as well as the starting point of each task, the resource(s) used, and the type of interaction that can occur between a human translator and a function or technological system. As a result, the model seeks to contribute to the development of professionals capable of making critical and efficient use of technologies in accordance with their specific professional contexts.

3 HCI Shifts In The Digitalised Translation Process By Chain-Of-Thought Prompting With ChatGPT

This section describes the tasks of each phase of the Tradumàtica model and explores the possibility of carrying them out with the assistance of an LLM such as ChatGPT. To accomplish this, we use the Chain-of-Thought Prompting (CoTP) model to send a reasoning-based action request to ChatGPT. CoTPs aim to provide the system with the conditions it needs to successfully complete the requested task.

The evaluation of the results takes place in three stages. The first stage of analysis allows us to determine whether the system correctly processed the natural language request. If this is the case, a paradigm shift in HCI in translation is underway: the translator will no longer manage the execution of each task through menus in graphical interfaces but will instead be able to limit himself/herself to ordering the task's execution.

The second stage of analysis allows us to determine whether the system is capable of responding to the request made. In this way, we can determine whether LLM outperforms any other existing translation technology solution.

The third level of analysis assesses how successful and appropriate the response obtained is in the context. This level of analysis allows us to determine whether the system is mature enough to be used professionally, or whether it needs to be improved.

As a result, given the CoTP for each task, we will examine:

- 1) Is ChatGPT capable of processing the request correctly?
- 2) Is ChatGPT capable of responding appropriately to the request format?
- 3) Is ChatGPT capable of successfully responding to the request?

The succeeding analysis describes the CoTPs for tasks in the Tradumàtica model's preparation, translation, and post-production phases. The phases of obtaining and returning the translation assignment have not been included because they are primarily dependent on the terms agreed upon by the client and the translation service provider. Simulated translation orders from English into European Spanish were used in all cases. To avoid relying on anecdotal evidence in our analysis, we created multiple prompts (no fewer than three) for each CoTP and documented the optimal response. CoTPs were also conducted in English or Spanish and in an English-Spanish translation context. All examples have been obtained through ChatGPT 3, and have also been compared with the results from ChatGPT 4 in order to provide the most up-to-date insight possible.

3.1 Preparation Phase And Tasks

This phase includes tasks as data analysis, linguistic analysis, reference material, pre-editing, and machine translation training. The aim of the phase is to obtain an overview of the project and decide what the translation resources and strategy will be. Preparing resources for translation involves configuring tools and resources and preparing external material. Section 4 includes a summary of how the LLM performed in each of the Tradumàtica model's tasks. This section delves deeper into the results obtained by using prompts to carry out some of the most common tasks in this phase.

Task: Data analysis. This task entails analysing data from the source files, such as word count, in order to create a time estimation and a pricing budget. Translators use a TEMS or text processors' features to run this analysis. Table 1 lists the CoTPs in terms of word count.

Table 1. CoTPs in task Word count.

CoTP	<p>1st Prompt: count words of a given text. 1st Answer: V3 apparently completed, but it failed. V4 successfully completed.</p> <p>2nd Prompt: count functional words and words with semantic meaning. 2nd Answer: V3 apparently completed, but it failed. V4 successfully completed.</p>
------	--

Results	The request seems to be correctly processed from the 1 st prompt, but version 3 does not really count words. Every time the system is asked to count words, it provides a different answer. However, version 4 does manage to achieve a correct result (see Appendix, table A-1).
---------	--

The system correctly processed the kind of requests it was asked for. However, despite providing a plausible answer (a number of words), the LLM system fails to execute the task by not running a proper word count. Nevertheless, the latest version of the language model is indeed capable of accurately computing the number of words on a sentence-by-sentence basis.

Table 2. CoTPs in task Pricing and time estimation.

CoTP	1 st Prompt: provide project pricing and execution time estimation (information provided: number of words, tools, resources and workforce involved). 1 st Answer: V3 failed – it provided pricing and time estimation considerations in response, but no calculation was provided. V4 successfully completed task.tin 2 nd Prompt: same request with more detailed information about the workforce and the time distribution between translation and review. 2 nd Answer: apparently inaccurately completed, but V3 failed – some figures were provided, but not as a result of a calculation, while V4 completed it successfully (see Appendix, Table A-2).
Results	The 1 st request failed. The 2 nd request based on more detailed information about each translator’s productivity threshold and the time distribution between initial translation and review obtained from V3 an appropriate although not precise answer, while the answer from V4 was more accurate.

Table 2 lists the CoTPs for pricing and time estimation. The system correctly processed the kind of requested it was asked for. Nevertheless, despite the provision of necessary information, V3 failed to execute the budget calculation. In contrast, the V4 was indeed capable of making a realistic task assignment and time management calculation, even taking into account criteria that were not provided in the prompt.

Task: Linguistic analysis. The purpose of these tasks is to gain a linguistic overview of the project as well as to delve into the text's and assignment's communicative situation in order to develop an appropriate translation strategy. To do so, translators use different features from TEMS and terminology management tools, as well as apply their translation theoretical and methodological knowledge and skills. Table 3 lists the CoPTs for terminology extraction.

Table 3. CoTPs in task Terminology extraction.

CoTP	1 st Prompt: extract terminology of a given text. 1 st Answer: completed. 2 nd Prompt: display terminology in a chart with a specified heading.
------	--

	2 nd Answer: completed.
	3 rd Prompt: display the same information in CSV format.
	3 rd Answer: completed.
	4 th Prompt: provide equivalents to the terms extracted.
	4 th Answer: inaccurately completed by V3 (equivalents are not accurate). Better results obtained with V4, even though with small errors.
	5 th Prompt: extract bilingual terminology from two parallel texts.
	5 th Answer: completed.
Results	A straightforward CoTP is enough to obtain a complete answer from the given texts (see Appendix, Table A-3)

The system ends completing successfully these tasks. It successfully performs monolingual term extraction, understanding “term” as relevant lexical items in the text rather than standardised terminology. The task that follows (Table 4) entails explicitly establishing the translation strategy, which reflects the decision-making framework for solving translation problems that a human translator performs, whether explicitly or implicitly, and that cannot be directly adopted by relying on a general-purpose MT system.

Table 4. CoTPs in task explicit setting of the translation strategy

CoTP	1 st Prompt: characterize the target reader of a given text. 1 st Answer: inaccurately completed by V3 (it provides the reader’s profile of interest, but not information about the register, tone, and other linguistic features), more accurately completed by V4. 2 nd Prompt: describe the prototypical target reader of a given text. 2 nd Answer: inaccurately completed by V3(it provides information about textual features, and infers the prototypical reader knowledge and interests), more accurately completed by V4. 3 rd Prompt: describe the reader based on a close classification of possible types of readers. 3 rd Answer: completed – it identifies the right type of reader of a given text by providing a scale of possible types.
Results	The request is processed from the first prompt, but the response is inaccurate and only meets expectations when the CoTP includes an explicit description of the types of readers to be considered (as a Likert scale) (see Appendix, Table A-4).

Task: Pre-editing. The aim of this task is to prepare the original text before the translation phase with (mainly) MT begins. The task consists of reducing or eliminating ambiguities and simplifying the syntax, with the goal of reducing the number of errors in the MT output. Translators use text processing and terminology management features. Table 5 lists the CoTPs regarding pre-editing.

Table 5. CoTPs in task Pre-editing

CoTP	1 st Prompt: define pre-editing. 1 st Answer: completed – it provides a full and correct description of it. 2 nd Prompt: pre-edit a given text with cultural references of the source audience and make it accessible to any foreign audience. 2 nd Answer: inaccurately completed by V3 – it removes cultural references and only explains those that are necessary for understanding the information in the source text. Accurately completed by V4 – it successfully identifies all cultural references that would be unintelligible to someone outside of that culture. It explains them (perhaps by providing more information than necessary) and even manages to adapt some of them. 3 rd Prompt: identify and extract cultural references in a given text that may not be globally understood. 3 rd Answer: inaccurately completed by V3 (it identifies many of the references of the text and defines them accurately. Accurately completed by V4. 4 th Prompt: translate the source text adapting all cultural references to the target audience. 4 th Answer: inaccurately completed by V3 (it provides a translation without cultural references and, consequently, with omissions), and by V4 (it provides unnecessary additions).
Results	In this case, the CoTP takes the form of a dialogue, because it requires explaining what the task is about before requesting an action. Depending on the prompt wording, the system seems to randomly apply the explicitation framework (see Appendix, Table A-4).

In these tasks, the system correctly processes the request and returns a response in the expected format, but response success is uneven.

Task: Translation. This is the core task of the translation process. The aim is to transfer the information on a text from a source language to a target language. The analysed subtasks of this task would be propagation of concordances, language resources and text editing. Translators use a TEMS, localizers or text processors to complete this task. Table 6 lists the CoTPs regarding translation.

Table 6. CoTPs in task Translation

Target-driven CoTP	1 st Prompt: translate a given text for a given target audience. 1 st Answer: completed (see Appendix, Table A-5).
Strategy-driven CoTP – example: transcreation	1 st Prompt: provide information regarding transcreation. 1 st Answer: completed. 2 nd Prompt: identify the proper transcreation techniques to translate a given text. 2 nd Answer: completed – it also provides the transcreation of the given text, even though it was not asked to (see Appendix, Table A-6 and Table A-7).

Results	The precise wording of the CoTP depends on the specific assignment and the translation strategy employed. A direct CoTP is enough for translations that follow a literal strategy (no content or form adaption to the target audience). However, when translation includes adaptation, transcreation or any other appropriate translation strategy, a previous dialogue is required to make the strategy explicit by ensuring the LLM receives the necessary details to create the most appropriate answer.
---------	---

For this task, the system appears to be far more versatile than a MT system. It allows for multiple translations from the same source by explicitly stating the translation strategy. However, because the amount of translation data analysed is limited, the conclusions about the quality of non-adapted translations (particularly in terms of accuracy) are only preliminary.

Task: Post-editing. It consists of a revision by a human translator of a text translated by an MT system. The translator edits the MT output segment by segment, accepting those without errors regarding the translation job. Table 7 lists CoTPs regarding post-editing.

Table 7. CoTPs in task Post-editing.

CoTP	1 st Prompt: define post-editing (PE). 1 st Answer: completed – it provides a full and accurate description of PE (see Appendix, Table A-8). 2 nd Prompt: post-edit the MT raw output from a given text. 2 nd Answer: almost accurately completed by V3 (it edits fluency errors, accuracy errors and mistranslations) and successfully completed by V4.
Results	The request is correctly processed on both the 1 st prompt and 2 nd prompt. The answer is appropriately executed but not completed due to model's capacity.

The task analysed consisted of post-editing a standard text in which the receiver of the source text and the target text share the same knowledge, so only the target text's fluency and accuracy in transmitting information needed to be ensured.

Task: Checking. It entails revising the text by the same translator who translated it to ensure that there are no errors in the translation or language errors regarding the source text and the assignment specifications. To complete this task, translators use a TEMS, terminology management tools, localizers, or text processors. Table 8 lists the CoTPs regarding the task language checking.

Table 8. CoTPs in task Language checking.

CoTP	1 st Prompt: identify terminology errors without providing a glossary.
------	---

	<p>1st Answer: V3 failed (it asserts that terminology is correct even when some terms are omitted), V4 completed it accurately.</p> <p>2nd Prompt: ensure a given glossary has been considered when translating a given text.</p> <p>2nd Answer: V3 failed (it claims that the terminology is correct, despite the fact that the translation contains several incorrect terms), V4 completed it accurately.</p> <p>3rd Prompt: convert the measurement units to a different metric system.</p> <p>3rd Answer: V3 failed (it only changes "Fahrenheit" to "Celsius," without recalculating the degrees), V4 completed it accurately.</p>
Results	The request appears to have been correctly processed; however, the newest version of the LLM assures that a glossary has been used consistently even though it has not. In the 3 rd Prompt, the model even succeeds to convert the measurements with calculations (see Appendix, Table A-9).

3.2 Post-Production Phase And Tasks

This is the last phase of the translation process. Its aim is to detect errors there might be present in the text and correct them. The translator that carries out these tasks is not necessarily the same as the one who fulfilled the initial translation.

Task: Linguistic post-production. This task involves the revision and the review of the text. Revision compares the target text to the source for accuracy and adequacy, while review monolingually evaluates stylistic and conceptual accuracy. This task is usually carried out using spellcheckers, TEMS, terminology managers, and many other DTP tools. Table 9 lists the CoTPs regarding revision and review.

Table 9. CoTPs in task Revision and Review.

CoTP	<p>1st Prompt: define translation review and revision.</p> <p>1st Answer: accurately completed – it provides a full definition that includes both translation revision and target text review.</p> <p>2nd Prompt: identify inconsistencies between source and target texts.</p> <p>2nd Prompt: not completed by V3 (it identifies explicit inconsistencies such as wrong dates or missing figures, but it does not identify accuracy errors or shifts of meaning), successfully completed by V4.</p> <p>3rd Prompt: identify and correct stylistic, tone, and register errors in the target text regarding the source text.</p> <p>3rd Answer: accurately completed.</p> <p>4th Prompt: confirm a given text is fluent in a target language.</p> <p>4th Answer: inaccurately completed by V3, successfully completed by V4.</p> <p>5th Prompt: edit properly a given text so that it answers to a different dialect or from a different ideological point of view.</p> <p>5th Answer: inaccurately completed.</p>
Results	From the first prompt, the revision and review request is correctly processed, but the task is incorrectly executed by the oldest version of the

LLM until the prompt includes an explicit description of the type of inconsistency that the system should consider (see Appendix, Table A-10).

The results obtained are uneven in the case of V3, and more accurate in the case of V4. The monolingual review (review of the target text) and rewriting of the target text from a different dialect or ideological point of view are both completed successfully. Nonetheless, it appears that the LLM is less capable of dealing successfully with cross-language revisions.

Task: Quality assessment. This task aims to ensure that quality standards are met. It entails ensuring that the project meets the quality requirements outlined in the project specifications. Table 10 lists the CoTPs regarding quality assessment actions such as applying regular expressions.

Table 10. CoTPs in task Regular expression.

CoTP	<p>1st Prompt: define what avoiding the abusive use of adverbs with the suffix <i>-mente</i> (<i>-ly</i> in English, one of the conditions provided in the specifications of the project) means.</p> <p>1st Answer: accurately completed – it provides a full and accurate description of this rule.</p> <p>2nd Prompt: edit the text as needed to fulfil this rule.</p> <p>2nd Answer: inaccurately completed by V3, accurately completed by V4.</p> <p>3rd Prompt: provide a regular expression to find this particular issue (adverbs ending in <i>-mente</i>) through another program.</p> <p>4th Prompt: inaccurately completed by V3 (it provides a regular expression with coding errors), accurately completed by V4.</p>
Results	The request is correctly processed from the 1 st prompt, but the final answer is not completed successfully. However, the model carries out the action the Regex syntax would be used for (see Appendix, Table A-11).

4 Results And Discussion

This article evaluates the execution capacity of the tasks as defined by the Tradumàtica model using ChatGPT (version 3) and its newest release (version 4). The study is based on the evaluation of results based on chain-of-thoughts prompts rather than just direct prompts. All evaluations are based on repeated CoTP and describe the best collected response. This section summarises in a concise manner the results obtained in all of the process's tasks, including some that were not described in the previous section. To that end, the tables below present the results of the three states of analysis described in section 3. Table 11 summarises the evaluation of CoTPs and ChatGPT 3 performance for Preparation phase tasks.

Table 11. Assessment of the CoTPs in preparation phase.

Task	Action	Stage 1 analysis	Stage 2 analysis	Stage 3 analysis
Data analysis	Word count	Yes	Yes (V4)	Successful (V4)
	Sentence division	Yes	Yes	Successful
	Pricing and time estimation	Yes	Yes	Successful (V4)
Linguistic analysis	Monolingual glossary creation	Yes	Yes	Successful
	Glossary in CSV format	Yes	Yes	Successful
Reference material	Bilingual Terminology extraction from parallel texts	Yes	Yes	Successful
	Monolingual terminology extraction and provision of equivalents	Yes	Yes	Successful (V4)
	Terminology definition	Yes	Yes	Inaccurate
	Chart and conceptual tree	Yes	Yes	Successful
	Search of referential material	Yes	Yes	Imprecise
Pre-editing	Identifying conceptual relations between terms	Yes	Yes	Successful
	Ambiguity identification	Yes	Yes	Successful (V4)
	Cultural reference explanation	Yes	Yes	Successful
	Translation with adaptation	Yes	Yes	Inaccurate

As shown in Table 11, one of the LLM's strengths is its ability to deal with terms and lexical units. The system even surpasses one of the greatest weaknesses of the previous version: its inconsistency in providing accurate results, especially when they rely on information not provided to the system via CoTP. Table 12, below, summarises the evaluation of CoTPs and ChatGPT 3 and 4 performance for Translation phase tasks.

Table 12. Assessment of the CoTPs in translation phase.

Task	Action	Stage 1 analysis	Stage 2 analysis	Stage 3 analysis
Translation	Target-driven translation:	Yes	Yes	Successful
	Strategy-driven translation:	Yes	Yes	Successful
Post-editing	Post-editing of a standard text	Yes	Yes	Successful
Checking	Identify terminology or conceptual error	Yes	No	Successful (V4)
	Edit terminology error	Yes	No	Successful (V4)

The outcomes of the tasks associated with this phase are very uneven. While it appears to perform well in translation and post-editing, it performs poorly in the initial revision (which is usually assumed by the translator).

Table 13. Assessment of the CoTPs in post-production phase.

Task	Action	Stage 1 analysis	Stage 2 analysis	Stage 3 analysis
Revision	Identify inconsistencies	Yes	No	Successful (V4)
	Correct inconsistency	Yes	No	Successful (V4)
Review	Context information	Yes	Yes	Successful
	Edit fluency errors	Yes	Yes	Successful
	Edit dialectical variant issues	Yes	Yes	Inaccurate
	Edit from an ideological perspective	Yes	Yes	Inaccurate
Quality assessment	Regex syntax	Yes	Yes	Successful (V4)

The results show that the system can correctly process natural language orders. This could revolutionise HCI in professional translation and training contexts. However, it has limitations when it comes to using its data for tasks other than translation, post-editing, and linguistic revision of the final text (by omitting or adding information). The findings indicate, on the one hand, the strengths and weaknesses of this type of LLM as a translation assistant. On the other hand, they suggest that some failed tasks by ChatGPT 3 may be satisfactorily resolved by ChatGPT 4, or even by utilising plugins or gateways with other programmes. However, LLMs cannot assume tasks such as determining the best translation strategy based on existing data without translation experts' prompts. Finally, the randomness and inconsistency of some of the results indicates that certain translation tasks requiring identical and precise information between the source and target text demonstrate that TA systems will remain professionally relevant.

References

1. Martín-Mor, A., Piqué, R., Sánchez-Gijón, P. *Tradumàtica: Tecnologies de la Traducció*. Eumo, (2016).
2. Holmes, J. S. *The name and nature of translation studies*. Amsterdam: University of Amsterdam Press (1975).
3. Hurtado Albir, A. *Researching translation competence by PACTE group*. John Benjamins Publishing. Amsterdam (2017).
4. Froeliger, N., Krause, A., Salmi, L. *Institutional translation–EMT Competence Framework and beyond. Institutional Translator Training*. Routledge. (2022).
5. Alcina, Amparo. "Translation technologies: Scope, tools and resources." *Target. International Journal of Translation Studies* 20.1 (2008): 79-102.
6. Olohan, M.: *Translators and translation technology: The dance of agency*. *Translation studies* 4(3), 342-357. (2011).
7. Pérez, Celia Rico. "La formación de traductores en Traducción Automática." *Revista Tradumàtica: tecnologies de la traducció* 15 (2017): 75-96.

8. Kenny, Dorothy. "Technology and translator training." *The Routledge handbook of translation and technology* (2019): 498-515.
9. Doherty, Stephen. "Translations| the impact of translation technologies on the process and product of translation." *International journal of communication* 10 (2016): 23.
10. Nirenburg, S.: *Tools for machine-Aided Translation: THE CMU TWS*. *Meta* 37(4), 709-720. <https://doi.org/10.7202/003739ar>. (1992).
11. Wilss, W.: *Basic Concepts of MT*. *Meta* 38(3), 403-413. <https://doi.org/10.7202/004608ar>, last accessed 2023/04/04. (1993).
12. Kay, Martin. "Machine translation." *American Journal of Computational Linguistics* 8.2 (1982): 74-78.
13. Massardo, I., van der Meer, J., O'Brien, S., Hollowod, F., Aranberri, N., Drescher, K.: *MT Post-Editing Guidelines*. TAUS, <https://info.taus.net/mt-post-editing-guidelines>, last accessed 2023/04/04. (2016).
14. Pérez-Ortiz, Juan Antonio, Mikel L. Forcada, and Felipe Sánchez-Martínez. "How neural machine translation works." *Machine translation for everyone: Empowering users in the age of artificial intelligence* 18 (2022): 141.
15. Tamkin, A., Brundage, M., Clark, J., Ganguli, D. *Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models*. *ArXiv*. /abs/2102.02503 last accessed 2023/04/04. (2021).
16. Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., Tao, D. *Towards Making the Most of ChatGPT for Machine Translation*. *ArXiv*. /abs/2303.13780 last accessed 2023/04/04. (2023).
17. Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., Awadalla, H. H. *How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation*. *ArXiv*. /abs/2302.09210 last accessed 2023/04/04. (2023).
18. Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., Tu, Z. *Document-Level Machine Translation with Large Language Models*. *ArXiv*. /abs/2304.02210 last accessed 2023/04/04. (2023).
19. Lu, Q., Qiu, B., Ding, L., Xie, L., Tao, D. *Error Analysis Prompting Enables Human-Like Translation Evaluation in Large Language Models: A Case Study on ChatGPT*. *ArXiv*. /abs/2303.13809 last accessed 2023/04/04. (2023).
20. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. *ArXiv*. /abs/2201.11903 last accessed 2023/04/04. (2022).
21. ISO. *ISO 17100 Translation services - Requirements for translation services*. Geneva: International Organization for Standardization. (2015).

Appendix

The analysis carried out in this article is rooted in examples of prompts between English and Spanish. This appendix includes some of the examples used for the analysis, featuring responses from ChatGPT 3 and 4. All prompts have been run more than once to avoid anecdotal and unrepresentative responses.

Table A-2. CoTPs in task Word count with ChatGPT 3 and ChatGPT 4

1 st	What is the total word count of the following text? "Machine-translated
<u>Prompt</u>	segments are increasingly included as fuzzy matches within the

	<p>translation-memory systems in the localisation workflow. This study presents preliminary results on the correlation between these two types of segments in terms of productivity and final quality. In order to test these variables, we set up an experiment with a group of eight professional translators using an on-line postediting tool and a statistical-base machine translation engine. The translators were asked to translate new, machine-translated and translation-memory segments from the 80-90 percent value using a post-editing tool without actually knowing the origin of each segment, and to complete a questionnaire. The findings suggest that translators have higher productivity and quality when using machine-translated output than when processing fuzzy matches from translation memories. Furthermore, translators' technical experience seems to have an impact on productivity but not on quality. Finally, we offer an overview of our current research."</p>
ChatGPT 3 – 1 st Answer	The total word count of the text is 123 words.
ChatGPT 4 – 1 st Answer	<p>The total word count of the given text is 184 words. Here's the breakdown:</p> <ul style="list-style-type: none"> - "Machine-translated segments are increasingly included as fuzzy matches within the translation-memory systems in the localisation workflow." - 17 words - "This study presents preliminary results on the correlation between these two types of segments in terms of productivity and final quality." - 19 words - "In order to test these variables, we set up an experiment with a group of eight professional translators using an on-line postediting tool and a statistical-base machine translation engine." - 27 words - "The translators were asked to translate new, machine-translated and translation-memory segments from the 80-90 percent value using a post-editing tool without actually knowing the origin of each segment, and to complete a questionnaire." - 33 words - "The findings suggest that translators have higher productivity and quality when using machine-translated output than when processing fuzzy matches from translation memories." - 21 words - "Furthermore, translators' technical experience seems to have an impact on productivity but not on quality." - 15 words - "Finally, we offer an overview of our current research." - 9 words <p>Adding all these together, 17 + 19 + 27 + 33 + 21 + 15 + 9, you get a total of 184 words.</p>

Table A-2. CoTPs in task Pricing and time estimation with ChatGPT 3 and ChatGPT 4

1 st Prompt	I have a team of 5 translators. Each translator translates an average of 100 words per hour. The project must pass final quality control. We
---------------------------	--

	<p>have the TAO memoQ tool, but we do not have previous translation memories for the project. The risk margin should be 3 hours. Give me the values so that all translators are involved in the translation process and the project is completed in two days.</p>
ChatGPT 3 – 1 st Answer	<p>In order for the team of 5 translators to complete the 20,000-word project in two days with a risk margin of 3 hours, the following steps can be taken:</p> <p>Establish a work plan: Divide the work into five equal parts, which means that each translator will have to translate approximately 4,000 words.</p> <p>Set deadlines: Set intermediate deadlines to ensure that the final deadlines are met. If working an 8-hour working day, each translator should deliver approximately 1,000 words every 8 hours.</p> <p>Use the TAO memoQ tool: Use the TAO memoQ tool to speed up the translation process and increase terminological consistency. Although there are no previous translation memories, the memoQ tool will allow translators to work more efficiently.</p> <p>Proofreading and editing: After each translator has finished his or her part, assign the work of proofreading and editing to another member of the translation team. Two revisions and final editing should be done before the work is handed over to the client.</p> <p>Final quality control: Perform a final quality control before delivering the work to the client to ensure that it meets the established quality standards.</p> <p>Considering that each translator can translate approximately 100 words per hour, to complete the 4,000 words assigned in two days (48 hours), each translator should work approximately 8 hours per day. Therefore, the work schedule would be as follows:</p> <p>Day 1: Each translator works for 8 hours and translates approximately 1,000 words.</p> <p>Day 2: Each translator works for 8 hours and translates approximately 1,000 words.</p> <p>Day 3: Each translator works for an additional 2 hours to complete the remaining 200 words and for the 3-hour risk margin.</p> <p>In summary, each translator must work 16 hours in total over two days to translate approximately 4,000 words each. In addition, intermediate deadlines must be set and the final translation must be proofread and edited before final quality control</p>
ChatGPT 4 – 1 st Answer	<p>If each translator translates an average of 100 words per hour and you have a team of 5 translators, the team's total translation capacity per hour is 500 words (5 translators * 100 words/hour).</p> <p>A two-day period consists of 48 hours. If we reserve 3 hours for a risk margin (for quality control and potential issues), the team will have 45 hours to work on the translation project. So, in total, the team can translate up to 22,500 words within this period (500 words/hour * 45 hours).</p>

	<p>To evenly distribute the workload and ensure all translators are involved, each translator would need to work 9 hours each day (45 hours total / 5 translators / 2 days) and translate 4,500 words over the course of two days (22,500 words total / 5 translators).</p> <p>So, for the project to be completed in two days with all translators involved and a 3-hour risk margin:</p> <ul style="list-style-type: none"> - The total project should be no more than 22,500 words - Each translator would need to work 9 hours each day for a total of 18 hours over two days - Each translator would need to translate a total of 4,500 words over the course of two days. 																								
2 nd Prompt:	<p>I have this staff for the team: One (junior translator), Two (junior translator), Three (junior senior translator), Four (senior translator), Five (senior translator). The project is 20,000, the tool to be used is SDL Trados. There are no translation memories to be used. The time we have is 7 days. 10% of the time must be allocated to the final revision, a global revision. Create a table with the assignment of tasks taking into account the position of each person.</p>																								
ChatGPT 3 – 2 nd Answer	<p>Taking into account the roles and experience of the team members, as well as the requirements of the project, we could assign the tasks as follows:</p> <table border="1"> <thead> <tr> <th>Translator</th> <th>Task</th> <th>Word count</th> <th>Estimated time</th> </tr> </thead> <tbody> <tr> <td>One (junior translator)</td> <td>Translation</td> <td>4,000</td> <td>2 days</td> </tr> <tr> <td>Two (junior translator)</td> <td>Translation</td> <td>4,000</td> <td>2 days</td> </tr> <tr> <td>Three (junior senior translator)</td> <td>Translation and proofreading</td> <td>6,000</td> <td>3 days</td> </tr> <tr> <td>Four (senior translator)</td> <td>Translation and proofreading</td> <td>4,000</td> <td>2 days</td> </tr> <tr> <td>Five (senior translator)</td> <td>Proofreading</td> <td>2,000</td> <td>1 day</td> </tr> </tbody> </table> <p>In this plan, junior translators would be responsible for part of the translation work, while senior translators would be responsible for the final revision and would also work on the translation of some parts of the project.</p> <p>The time allocated for each task is an estimate based on the number of words and the experience of each translator. In addition, 10% of the total time would be allocated to an overall final review, where all team members could work together to ensure the quality of the project.</p> <p>It is important to note that these times are only estimates and may vary depending on the complexity of the project and the individual skills of each team member.</p>	Translator	Task	Word count	Estimated time	One (junior translator)	Translation	4,000	2 days	Two (junior translator)	Translation	4,000	2 days	Three (junior senior translator)	Translation and proofreading	6,000	3 days	Four (senior translator)	Translation and proofreading	4,000	2 days	Five (senior translator)	Proofreading	2,000	1 day
Translator	Task	Word count	Estimated time																						
One (junior translator)	Translation	4,000	2 days																						
Two (junior translator)	Translation	4,000	2 days																						
Three (junior senior translator)	Translation and proofreading	6,000	3 days																						
Four (senior translator)	Translation and proofreading	4,000	2 days																						
Five (senior translator)	Proofreading	2,000	1 day																						
ChatGPT 4 – 2 nd Answer	<p>To construct a table for task assignment, I will assume that senior translators are able to translate 15% more words per hour than junior translators due to their experience and efficiency, and that a junior senior translator translates 7.5% more words per hour than a junior translator (halfway between a junior and a senior translator). Let's also</p>																								

assume that a junior translator translates 100 words per hour (this is an average and real-life values can be different). Here are the translation rates I'm using based on the above assumptions:

- Junior translator: 100 words/hour
- Junior senior translator: 107.5 words/hour
- Senior translator: 115 words/hour

The total project size is 20,000 words and we have 7 days to complete it, with 10% of the time allocated for a final global revision. This means we have 6.3 days for translation (90% of 7 days) and 0.7 days for the final revision.

Now, let's divide the translation tasks:

Translator	Position	Words/Hour	Hours/Day	Days	Total Words
One	Junior translator	100	7	6.3	4410
Two	Junior translator	100	7	6.3	4410
Three	Junior senior translator	107.5	7	6.3	4745
Four	Senior translator	115	7	6.3	5080
Five	Senior translator	115	7	6.3	5080
Total					23725

The total word count we have is slightly above 20,000 words, this means that we are safely within our limit even if translators are a little slower than expected. The remaining words (~1725) provide a buffer and can be divided among translators as needed.

For the final revision, which requires 10% of the time (or 0.7 days), you may want to assign this to the senior translators (Four and Five) due to their greater experience, but it could also be shared amongst the team depending on how your team works best.

This table provides an outline and can be modified as per the actual productivity of the translators, their availability, and other project-specific factors. The main aim is to distribute the work according to the translators' experience levels to ensure efficiency while also ensuring all translators are involved.

Table A-3. CoTPs in task Terminology extraction with ChatGPT 3 and ChatGPT 4

1 st Prompt	Extract terminology of the following text: “A fin de limitar la carga administrativa, los Estados miembros deben poder transmitir datos estadísticos y el importe del recurso propio en un único estado. Los datos sobre la generación y el reciclado de residuos de envases de plástico constituyen la base para calcular las contribuciones nacionales al presupuesto general de la Unión. Por ello, es necesario que se refuerce la
------------------------	--

	comparabilidad, la fiabilidad y la exhaustividad de esos datos. A fin de garantizar la comparabilidad, fiabilidad y exhaustividad de los datos compartidos entre los Estados miembros, conviene establecer normas detalladas sobre los datos que han de figurar en el estado que debe facilitarse a la Comisión”.
ChatGPT 3 – 1 st Answer	<ul style="list-style-type: none"> ○ Estados miembros ○ carga administrativa ○ datos estadísticos ○ recurso propio ○ generación ○ reciclado ○ residuos de envases de plástico ○ contribuciones nacionales ○ presupuesto general de la Unión ○ comparabilidad ○ fiabilidad ○ exhaustividad ○ normas detalladas ○ Comisión
ChatGPT 4 – 1 st Answer	<p>The following terms from the text can be seen as specific terminology:</p> <ol style="list-style-type: none"> 1. Carga administrativa 2. Estados miembros 3. Datos estadísticos 4. Importe del recurso propio 5. Generación de residuos 6. Reciclado de residuos 7. Residuos de envases de plástico 8. Contribuciones nacionales 9. Presupuesto general de la Unión 10. Comparabilidad 11. Fiabilidad 12. Exhaustividad 13. Normas detalladas 14. Comisión <p>Each term corresponds to various concepts within the field of international administrative and environmental law, especially as it pertains to the European Union.</p>
2 nd Prompt	Extract the terminology on these texts and display it on a chart with the headings “Término en español” and “Término en inglés”. Perform the bilingual extraction of terminology from these two texts: “A fin de limitar la carga administrativa, los Estados miembros deben poder transmitir datos estadísticos y el importe del recurso propio en un

único estado. Los datos sobre la generación y el reciclado de residuos de envases de plástico constituyen la base para calcular las contribuciones nacionales al presupuesto general de la Unión. Por ello, es necesario que se refuerce la comparabilidad, la fiabilidad y la exhaustividad de esos datos. A fin de garantizar la comparabilidad, fiabilidad y exhaustividad de los datos compartidos entre los Estados miembros, conviene establecer normas detalladas sobre los datos que han de figurar en el estado que debe facilitarse a la Comisión.”

“In order to limit the administrative burden, Member States should be able to transmit statistical data and the amount of the own resource in a single statement. Data on plastic packaging waste generation and recycling constitute the basis for calculating the national contributions to the general budget of the Union. Therefore, it is necessary to reinforce the comparability, reliability and exhaustiveness of those data. In order to guarantee the comparability, reliability and exhaustiveness of data between the Member States, it is appropriate to lay down detailed rules concerning the data to be contained in the statement to be provided to the Commission.”

ChatGPT
3 –2nd
Answer

Término en español	Término en inglés
carga administrativa	administrative burden
Estados miembros	Member States
datos estadísticos	statistical data
recurso propio	own resource
único estado	single statement
generación de residuos de envases de plástico	plastic packaging waste generation
reciclado de residuos de envases de plástico	plastic packaging waste recycling
contribuciones nacionales	national contributions
presupuesto general de la Unión	general budget of the Union
comparabilidad	comparability
fiabilidad	reliability

	exhaustividad	exhaustiveness
	normas detalladas	detailed rules

ChatGPT
4 – 2nd
Answer

Término en español	Término en inglés
Carga administrativa	Administrative burden
Estados miembros	Member States
Datos estadísticos	Statistical data
Importe del recurso propio	Amount of the own reso
Generación de residuos	Waste generation
Reciclado de residuos	Waste recycling
Residuos de envases de plástico	Plastic packaging waste
Contribuciones nacionales	National contributions
Presupuesto general de la Unión	General budget of the U
Comparabilidad	Comparability
Fiabilidad	Reliability
Exhaustividad	Exhaustiveness
Normas detalladas	Detailed rules
Comisión	Commission

3rd Display the same information in CSV format.

Prompt

ChatGPT
3 – 3rd
Answer

Término en español,Término en inglés
Carga administrativa,Administrative burden
Estados miembros,Member States
Datos estadísticos,Statistical data
Recurso propio,Own resource
Generación,Generation
Reciclado,Recycling
Residuos,Waste
Envases de plástico,Plastic packaging
Contribuciones nacionales,National contributions
Presupuesto general,General budget
Unión,Union

	Comparabilidad, Comparability Fiabilidad, Reliability Exhaustividad, Exhaustiveness Normas detalladas, Detailed rules Comisión, Commission
ChatGPT 4 – 3 rd Answer	"Término en español", "Término en inglés" "Carga administrativa", "Administrative burden" "Estados miembros", "Member States" "Datos estadísticos", "Statistical data" "Importe del recurso propio", "Amount of the own resource" "Generación de residuos", "Waste generation" "Reciclado de residuos", "Waste recycling" "Residuos de envases de plástico", "Plastic packaging waste" "Contribuciones nacionales", "National contributions" "Presupuesto general de la Unión", "General budget of the Union" "Comparabilidad", "Comparability" "Fiabilidad", "Reliability" "Exhaustividad", "Exhaustiveness" "Normas detalladas", "Detailed rules" "Comisión", "Commission"

Table A-4. CoTPs in task Pre-editing with ChatGPT 3 and ChatGPT 4

1 st Prompt	What is pre-editing in translation?
ChatGPT 3 – 1 st Answer	The term "pre-editing" in translation refers to the process of reviewing and editing a translated text before it is sent to the client or published. In pre-editing, an expert editor or proof reader checks the translation for errors in spelling, grammar, punctuation, coherence and cohesion, as well as to ensure that the tone, style and terminology are accurate and appropriate for the target audience. Pre-editing is an important part of the translation process because it helps ensure that the translated text is of high quality and meets the client's requirements and expectations. Pre-editing also helps ensure that the translated text is clear, consistent, and easy to understand for readers who speak the target language.
ChatGPT 4 – 1 st Answer	Pre-editing in the context of translation typically refers to the process of preparing a text for machine translation. This involves cleaning and standardizing the text to improve the quality of the machine translation output. Pre-editing can include: 1. Correcting spelling and grammar errors in the source text. 2. Simplifying complex sentences or breaking them down into shorter, more straightforward sentences. 3. Removing or explaining idiomatic expressions that may not translate well. 4. Standardizing terminology, especially in technical texts.

	<p>5. Formatting the text consistently, including things like dates, times, and numbers.</p> <p>The goal of pre-editing is to make the source text as clear and unambiguous as possible, as this can help the machine translation system to produce a more accurate and readable translation. After the machine translation, post-editing is often required to polish the translated text and make any necessary corrections or improvements. This process is crucial in ensuring high-quality translations when using machine translation systems.</p>
2 nd Prompt	<p>Pre-edit this text so that the cultural references are not a problema for the understanding of the text by a reader who des not speak spanish nor knows the spanish culture. “Algunos sitúan el inicio de la temporada navideña en el Black Friday, otros, en el puente del Pilar. El consenso, sin embargo, se sitúa en una fecha concreta que combina muchos de los ingredientes de la época: el 22 de diciembre y su Lotería de Navidad. Encuentro, gasto, una canción repetitiva —"miiiiil eeeeu-roooooo"— que se instala en el cerebro durante horas o días, mucha ilusión velada y la esperanza de que una lluvia de dinero caiga cerca. Esta ilusión siempre existe, aunque la probabilidad juega en contra.</p> <p>La estadística es clara. Hay cien mil números, del 00.000 al 99.999. Solo una de estas cifras resultará agraciada con el Gordo de cuatro millones de euros (400.000 para cada décimo de los 10 que forman parte de un número entero). Uno de cien mil, la misma proporción de que toque un segundo o un tercer premio (agraciados con 125.000 y 50.000 cada cupón respectivamente). Esta es una probabilidad muy parecida a acertar al señalar con los ojos cerrados a una persona concreta en un Camp Nou con sus 99.354 localidades ocupadas. O hacerlo con un pelo concreto de los cien mil que tenemos de media los seres humanos en la cabeza.”</p> <p>pre-edit a given text with cultural references of the source audience and make it accessible to any foreign audience.</p>
ChatGPT 3 – 2 nd Answer	<p>Algunos sitúan el inicio de la temporada navideña en el Black Friday, un día de compras masivas que se celebra después del Día de Acción de Gracias en Estados Unidos. Otros, lo sitúan en una fecha en concreto: el 22 de diciembre, cuando se celebra la Lotería de Navidad, un sorteo de lotería muy popular en España. En este día, muchas personas tienen la esperanza de ganar grandes premios, aunque las posibilidades son muy bajas. En el sorteo de la Lotería de Navidad, hay cien mil números posibles, pero solo uno de ellos recibe el premio más grande, conocido como el "Gordo". También hay premios para otros números, pero las probabilidades son similares a encontrar una persona con los ojos cerrados en un estadio lleno de gente o encontrar un cabello específico entre los muchos que tiene una persona. Aunque la Lotería de Navidad es muy importante para muchos españoles, es posible que no sea un evento conocido fuera de España.</p>

ChatGPT 4 – 2 nd Answer	<p>Sure, here is a pre-edited version of the text where the cultural references have been explained or generalized for a reader who may not be familiar with Spanish culture:</p> <p>"Algunos marcan el inicio de la temporada navideña en el día después del Día de Acción de Gracias, conocido internacionalmente como Black Friday, otros, a mediados de octubre. El consenso, sin embargo, generalmente se sitúa en una fecha específica que combina muchos de los elementos de esta época: el 22 de diciembre y su sorteo anual de Navidad, similar a las loterías de fin de año que se celebran en muchos países. Este evento fomenta reuniones sociales, implica gastos y se asocia con una canción pegadiza de temática lotería que permanece en la mente durante horas o incluso días. Es un tiempo lleno de anticipación y la esperanza de recibir repentinamente una gran suma de dinero. Esta esperanza siempre está ahí, aunque las probabilidades estén en contra.</p> <p>Las estadísticas son claras. Hay cien mil números posibles, del 00.000 al 99.999. Solo uno de estos números ganará el premio principal de cuatro millones de euros (400.000 euros para cada décimo de un billete completo). La posibilidad de ganar el premio principal, o los premios segundo o tercero (que ofrecen 125.000 y 50.000 euros respectivamente por billete), es de una en cien mil. Esto es similar a seleccionar aleatoriamente a una persona específica de una multitud en un estadio de fútbol completamente ocupado, como cualquier estadio grande con capacidad para 99.354 espectadores, o seleccionar un solo cabello de entre los cien mil pelos que tiene en promedio un ser humano en la cabeza."</p>
3 rd Prompt	Which are the elements a reader from another country who knows nothing about the Spanish culture will not understand?
ChatGPT 3 – 3 rd Answer	<p>The text contains specific references to Spanish culture and traditions that might not be understood by a reader from another country without prior knowledge. Some examples are:</p> <ul style="list-style-type: none"> o Blak Friday: The term "Blak Friday" is a misspelling of "Black Friday", the day of mass shopping after Thanksgiving in the United States. Although Black Friday has also been adopted in other countries, the term "Blak" can be confusing for a foreign reader. o Puente del Pilar: The "Puente del Pilar" refers to the long weekend around 12 October, the day of Spain's National Day and also the Day of the Virgen del Pilar, the patron saint of Spain. This term and its meaning may not be familiar to a reader from another country. o Lotería de Navidad: The Lotería de Navidad is a very popular lottery draw in Spain that is held every 22 December. The text mentions that this date is very important for many Spaniards as the start of the Christmas season, but a foreign reader may not be familiar with this tradition. o Gordo, segundo y tercer premio: In the Lotería de Navidad, prizes are distributed in several levels, and the three main ones are the Gordo (the biggest prize), the second prize and the third prize. These terms

	<p>and their importance in the Christmas Lottery draw may not be understood by a foreign reader.</p> <p>In general, the text contains many specific references to Spanish culture and traditions, so a foreign reader may need a greater understanding of these references to fully understand the text.</p>
ChatGPT 4 – 3 rd Answer	<p>In the original Spanish text, the following cultural references may be difficult to understand for a reader from another country unfamiliar with Spanish culture:</p> <ol style="list-style-type: none"> 1. "Black Friday": While Black Friday is well-known in the United States and has gained popularity in many other countries due to globalization, not all countries observe this shopping holiday. 2. "Puente del Pilar": Literally translates to "The Pillar's Bridge". This is a reference to the Spanish holiday that takes place on October 12th, the Feast of Our Lady of the Pillar. In Spain, a "puente" (bridge) refers to a long weekend or a stretch of days off made by taking additional leave around public holidays. This concept may not be understood in countries where this is not common practice. 3. "Lotería de Navidad": The Christmas Lottery in Spain is a major event with a long history, dating back to 1812. It has a unique format that may not be familiar to people outside of Spain. 4. "miiii eeeurooooooos": This is a reference to a jingle or catchphrase associated with the Christmas Lottery. People unfamiliar with this tradition would not understand the significance. 5. "Gordo": This term literally means "fat one" and is used to refer to the grand prize in the Spanish Christmas Lottery. 6. "decimo": A "decimo" in the context of the Spanish Lottery is a tenth of a whole ticket. This is unique to the format of the Spanish lottery where whole tickets are often too expensive for one person to buy, so they are split into 'decimos'. 7. "Camp Nou": This is a football stadium in Barcelona, Spain, and is the home stadium of FC Barcelona. The comparison of the probability of winning the lottery to picking out a specific person in the packed Camp Nou stadium assumes the reader knows about this specific stadium and its capacity. 8. "un segundo o un tercer premio": The Spanish Christmas Lottery has a unique structure with several prize levels, which may not be familiar to those outside of Spain. <p>These terms and cultural references are very specific to Spain and might not be well-known to people from other countries or cultures. In a pre-edited version, these terms would need to be replaced or explained.</p>
4 th Prompt	Translate the pre-edited text into English.
ChatGPT 3 – 4 th Answer:	Some people consider the start of the Christmas season to be Black Friday, a day of massive shopping that takes place after Thanksgiving in the United States. Others place it on a specific date: December 22, when the Christmas Lottery is held, a very popular lottery in Spain.

	<p>On this day, many people hope to win big prizes, although the chances are very low. In the Christmas Lottery, there are one hundred thousand possible numbers, but only one of them receives the biggest prize, known as the "Gordo". There are also prizes for other numbers, but the odds are similar to finding a specific person with your eyes closed in a stadium full of people or finding a specific hair among the many that a person has. Although the Christmas Lottery is very important for many Spaniards, it may not be a well-known event outside of Spain.</p>
ChatGPT 4 – 4 th Answer	<p>Here is the English translation of the pre-edited Spanish text: "Some mark the start of the Christmas season on the day after Thanksgiving, known internationally as Black Friday, others, in mid-October. However, consensus generally places it on a specific date that combines many elements of this time: December 22nd, and its annual Christmas draw, similar to the end-of-year lotteries celebrated in many countries. This event promotes social gatherings, involves spending, and is associated with a catchy lottery-themed song that lingers in the mind for hours or even days. It's a time filled with anticipation and the hope of suddenly receiving a large sum of money. This hope is always there, even though the odds are against it. The statistics are clear. There are one hundred thousand possible numbers, from 00,000 to 99,999. Only one of these numbers will win the main prize of four million euros (400,000 euros for each tenth of a full ticket). The chance of winning the main prize, or the second or third prizes (which offer 125,000 and 50,000 euros respectively per ticket), is one in a hundred thousand. This is similar to randomly picking a specific person from a crowd in a fully occupied football stadium, like any large stadium with a capacity of 99,354 spectators, or picking a single hair from the hundred thousand hairs that an average human being has on their head."</p>

Table A-5. CoTPs in task Translation (target-driven) with ChatGPT 3 and ChatGPT 4

1 st Prompt	<p>Translate the following text into English taking into account the target readers are children: "Desde la publicación de su descubrimiento en la famosa Formación Hell Creek -HCF- en 1905, el tiranosaurio rex ha sido el foco de un intenso interés científico y público que persiste hasta nuestros días. Las numerosas hipótesis sobre la biología y el comportamiento de T. rex con las que contamos hasta el momento son el resultado de décadas de investigación centradas principalmente en la morfología esquelética y la biomecánica de estos temibles animales. Sin embargo, hasta hace muy poco tiempo, apenas los últimos 15 años, no se habían aplicado otras técnicas como la histología ósea para investigar algunos aspectos inaccesibles a partir de exámenes generales</p>
------------------------	---

	y hasta el momento desconocidos de la vida del T. rex. Algunos de estos aspectos abordan cuestiones que abarcan desde el desarrollo embrionario, pasando por la tasa de crecimiento o la madurez esquelética, hasta la madurez sexual.”
ChatGPT 3 – 1 st Answer	Since the publication of its discovery in the famous Hell Creek Formation -HCF- in 1905, the tyrannosaurus rex has been the focus of intense scientific and public interest that persists to this day. The numerous hypotheses about the biology and behavior of T. rex that we have to date are the result of decades of research focused mainly on the skeletal morphology and biomechanics of these fearsome animals. However, until very recently, just the last 15 years, other techniques such as bone histology had not been applied to investigate some aspects that were inaccessible from general exams and up until now unknown about the life of T. rex. Some of these aspects address questions ranging from embryonic development, through growth rate or skeletal maturity, to sexual maturity.
ChatGPT 4 – 1 st Answer	Here's a child-friendly translation of your text into English: "Ever since its discovery was published in the famous Hell Creek Formation - HCF - in 1905, the Tyrannosaurus Rex, or T-Rex for short, has been the center of lots of excitement and interest, both for scientists and dinosaur fans like us. We've made a lot of guesses about how the T-Rex lived and behaved, and these guesses are based on years and years of study focused mainly on what its bones look like and how it might have moved around. However, until very recently - just the last 15 years - we haven't used other methods like studying bone tissues to learn more about parts of the T-Rex's life that we can't figure out just by looking at its bones. Some of these things include how T-Rex babies developed before they were born, how fast they grew, when they became full-grown, and when they were old enough to have babies of their own."

Table A-6. CoTPs in task explicit setting of the translation strategy ChatGPT 3 and ChatGPT 4

1 st Prompt	Which are the characteristics of the target reader of the following text? “Do you know how we know that climate change is really happening? Climate change is often presented as a prediction made by complicated computer models, but the scientific basis for climate change is much broader, and models are only one part of it (and, for what it's worth, they're surprisingly accurate).
---------------------------	---

For over a century, scientists have understood the basic physics behind why greenhouse gases like carbon dioxide cause warming. These gases make up just a small fraction of the atmosphere, but they have a lot of control over Earth's climate by trapping some of the planet's heat before it escapes into space. This greenhouse effect is important: it's why a planet so far from the sun has liquid water and life!

But during the Industrial Revolution, people started burning coal and other fossil fuels to power factories, smelters, and steam engines, which added more greenhouse gases to the atmosphere. Since then, human activities have been heating the planet.

We know this is true thanks to a ton of evidence that starts with temperature measurements taken at weather stations and on ships starting in the mid-1800s. Later, scientists began tracking surface temperatures with satellites and looking for clues about climate change in geological records. All of this data tells the same story: Earth is getting hotter.

Average global temperatures have increased by 2.2 degrees Fahrenheit, or 1.2 degrees Celsius, since 1880, with the greatest change happening in the late 20th century. Land areas have warmed more than the sea surface, and the Arctic has warmed the most – by more than 4 degrees Fahrenheit just since the 1960s. Temperature extremes have also shifted. In the US, daily record highs now outnumber record lows two-to-one.

This warming is unprecedented in recent geological history.”

ChatGPT 3 – 1 st Answer	Characteristics of the target reader of this text could include: <ul style="list-style-type: none">o Interest in climate change: It is important that the recipient has an interest in climate change and in obtaining information about it.o General level of knowledge: It is important that the receiver has a general level of knowledge about climate change and related scientific and technical concepts.o Reading and comprehension skills: It is important that the recipient has good reading and comprehension skills to be able to understand the content of the text.o Availability of time: It is important that the receiver has time to read and analyse the text thoroughly. Openness to consider different perspectives: It is important that the receiver is willing to consider different perspectives and arguments on climate change.
ChatGPT 4 – 1 st Answer	The target reader of the provided text seems to possess the following characteristics: <ol style="list-style-type: none">1. General Audience: The language used in the text is straightforward, clear, and not overly academic, suggesting that it is intended for

	<p>a general audience rather than specialists. The text avoids technical jargon, indicating that it is likely aimed at the lay public.</p> <p>2. Interest in Climate Change: Given the subject matter, it's clear the text is aimed at readers who have an interest in understanding the concept of climate change, its causes, and its effects.</p> <p>3. Educational Purpose: The text is explanatory and provides basic information on a complex subject, suggesting it's for educational purposes. Therefore, it could be targeting students or individuals seeking to enhance their knowledge about the environment and global warming.</p> <p>4. Non-scientific Background: The text presents scientific concepts like the greenhouse effect and global warming in an accessible way, indicating it's aimed at individuals who do not necessarily have a scientific background.</p> <p>5. Interest in Environmental Issues: The text's emphasis on historical data and the impacts of human activity suggest a target reader who is concerned with or interested in environmental issues and sustainability.</p> <p>6. Language Proficiency: The reader is expected to be proficient in English, given that the text uses idiomatic expressions like "a ton of evidence", metaphorical language like "planet's heat before it escapes into space", and higher-level vocabulary words like "Industrial Revolution", "geological records", and "unprecedented".</p>
2 nd Prompt	What is the type of reader of the text I have provided above?
ChatGPT 3 – 2 nd Answer	The text you provided me with above is a scientific-informative text that focuses on climate change. The language used is technical and precise and draws on a wealth of evidence and data to explain the scientific processes behind climate change. The text also includes a detailed description of the tools and methods used to measure and analyse climate, as well as the observed consequences of climate change. The typical reader would be a person with a background in environmental science, physics, chemistry, geology, statistics and mathematics.
ChatGPT 4 – 2 nd Answer	The type of reader for the provided text is a non-specialist, layperson, or general reader who has an interest in understanding climate change, global warming, and related environmental issues. This reader likely does not have a specialized background in science or environmental studies, but has a basic interest in these topics. The reader is likely to be someone seeking to enhance their general knowledge or understanding of these issues, perhaps for personal interest, education, or to make more informed decisions about sustainable practices. They are proficient in English and comfortable with reading content that uses some advanced vocabulary and idiomatic expressions.
3 rd Prompt	For which type of reader do you think this text is most suitable: A) Physics student, medium level of knowledge on the subject, use of the text for a university assignment. B) Scientist, specialist in the field of

	climate change, using the text for leisure reading. C) Non-scientist, interested in the topic and reads the text to learn more about it. D) Scientist from another field outside climate change and climate, reads the text in a scientific journal. E) Average reader, basic knowledge, basic education.
ChatGPT 3 – 3 rd Answer	I think the text is more suitable for audience B) Scientist, specialist in the field of climate change, use of the text for leisure reading. The text contains precise, technical language and draws on a wealth of evidence and data to explain the scientific processes behind climate change, so a scientist specialising in this field would be able to understand and appreciate the content of the text. Although it could be interesting for the audience A) Physics student, medium level of knowledge on the topic, using the text for a university assignment, as the text provides a lot of details and evidence on climate change, but would be difficult to understand for an audience C) Non-scientist, interested in the topic and reading the text to learn more about it. D) Scientist from another field outside climate change and climate, reads the text in a scientific journal, as the text is specific to the topic of climate change and does not relate to other scientific areas, and E) Average reader, basic knowledge, basic training, as the text is technical and requires prior and specific knowledge in the areas mentioned above.
ChatGPT 4 – 3 rd Answer	Based on the language, depth of the content, and the way the information is presented, the text seems most suitable for: C) Non-scientist, interested in the topic and reads the text to learn more about it. This text provides a broad overview of the topic and is written in a way that is accessible to a general audience. The complexity of the text is appropriate for a non-scientist who is interested in climate change and is seeking to learn more about it. The text might be too basic for a physics student doing a university assignment (option A), a scientist who is a specialist in the field of climate change reading for leisure (option B), or a scientist from another field reading in a scientific journal (option D). For these readers, the content might lack the depth, complexity, or specificity that they would typically look for. Option E) Average reader, basic knowledge, basic education could also be a potential reader type, but the text does include some advanced vocabulary and concepts that might be challenging for a reader with only basic education.

Table A-7. CoTPs in task Translation – Strategy-driven with ChatGPT 3 and ChatGPT 4

1 st Prompt	What is transcreation?
------------------------	------------------------

ChatGPT 3 – 1 st Answer	<p>Transcreation is a term used in marketing and advertising to refer to the adaptation of an advertising message, slogan, campaign or any other type of creative content from one language to another, maintaining its meaning, impact and effectiveness in the new language and in the culture of the target audience.</p> <p>Unlike translation, which focuses on conveying the literal meaning of a text from one language to another, transcreation involves adapting the original message to the culture, values, manner of expression and tone of voice of the target audience in the new language. This implies that transcreation is not limited to simple translation, but also involves the recreation of the creative elements, such as metaphors, puns, images and feelings associated with the original message.</p> <p>Transcreation is especially important in cases where the creative content has a strong emotional impact on the target audience and its effectiveness is highly dependent on the cultural understanding of the message. Examples of content that can be transcreated include television commercials, billboards, websites, brochures, e-mail messages and other types of creative content used in marketing and advertising.</p>
ChatGPT 4 – 1 st Answer	<p>Transcreation is a process used in the field of translation that involves not just translating text from one language to another, but also creating new content in the target language that conveys the same message and evokes the same emotions as the original text. The term is a blend of 'translation' and 'creation'.</p> <p>The goal of transcreation is to maintain the intent, tone, and emotion of the original message while making it culturally appropriate for the target audience. This is especially important in areas such as marketing and advertising, where it's crucial to engage and connect with the audience on a cultural and emotional level. It often involves creative writing skills and a deep understanding of both the source and target cultures.</p> <p>For example, a marketing slogan that is a play on words in English may not make sense if it's directly translated into another language. Instead, a transcreator would come up with a new slogan in the target language that conveys the same message and has the same emotional impact.</p> <p>Transcreation goes beyond literal translation and requires more creative effort, making it a separate field of expertise within the broader translation industry.</p>
2 nd Prompt	<p>What transcreation tactics would you use for the following text? It should be in English for a UK audience: "Tú lo quieres fácil. Nosotros lo tenemos a un clic. Los precios que buscas, los tenemos nosotros. Aprovecha ya las ofertas exclusivas de nuestra web."</p>
ChatGPT 3 – 2 nd Answer	<p>To transcreate this text into English for a UK audience, the following transcreation tactics could be used:</p> <p>Adaptation of idioms and tone of voice: "Tú lo quieres fácil" could be translated into English as "You want it easy", and "Aprovecha ya" could be translated as "Get it now". In addition, for English audiences, a more formal or informal tone of voice could be used, depending on the target audience.</p>

	<p>Use of local words and phrases: The text could be adapted using specific British English words and phrases to make it more relevant to the UK audience. For example, you could use "You're after an easy way" instead of "You want it easy" and "Take advantage of our website's exclusive deals" instead of "Aprovecha ya las ofertas exclusivas de nuestra web".</p> <p>Consideration of culture and values: To adapt the message to the culture and values of the audience in England, the tone and style of the text could be adjusted. For example, a more humorous or sarcastic tone could be used to make the message more appealing to the UK audience. Using these tactics, the text could be transcreated into English as follows: "You're after an easy way. We've got it just a click away. The prices you're looking for? We've got them. Take advantage of our website's exclusive deals now".</p>
ChatGPT 4 – 2 nd Answer	<p>To transcreate this text for a UK audience, it would be important to consider not just literal translation, but also cultural nuances and idioms that would resonate with this audience. Here's a possible transcreated version:</p> <p>"You want it simple. We've got it one click away. The prices you're after, we've got them. Grab exclusive online deals now."</p> <p>Transcreation tactics used here:</p> <ol style="list-style-type: none"> Adaptation of Idioms: "Tú lo quieres fácil" has been transcreated as "You want it simple." The concept of wanting things to be easy or hassle-free is universal, but the expression has been adapted to fit more naturally into English. Adaptation of Call-to-Action phrases: "Aprovecha ya las ofertas exclusivas de nuestra web" has been translated and adapted into a more culturally fitting call to action: "Grab exclusive online deals now". "Grab" is a more casual and engaging term commonly used in UK English for such prompts. Keeping Meaning and Tone: The overall tone of the text, which is casual and customer-oriented, has been maintained in the transcreated version. This ensures the message continues to appeal to the target audience in the way the original text intended.

Table A-8. CoTPs in task Post-editing with ChatGPT 3 and ChatGPT 4

1 st Prompt	What is post-editing (PE)?
ChatGPT 3 – 1 st Answer	<p>Post-editing" (also known as "post-editing") refers to the process of reviewing and editing machine-generated text, such as a machine translation system. In post-editing, a human reviews and corrects the machine-generated text to ensure that it is consistent, accurate and fit for purpose.</p> <p>Post-editing is common in the translation industry, where it is used to improve the quality of machine translation. It is also used in other</p>

	fields where texts are automatically generated, such as in automatic speech-to-text transcription, summary generation and automatic report writing. Post-editing can help improve the efficiency and accuracy of the automated text generation process, allowing humans to work on more complex and creative tasks.
ChatGPT 4 – 1 st Answer	<p>Post-editing (PE) is a process in the field of translation where human translators review, correct, and improve machine-generated translations. This process is particularly prevalent in the context of Machine Translation (MT), where a text is initially translated by an automated system and then polished by a human post-editor.</p> <p>There are generally two levels of post-editing:</p> <ol style="list-style-type: none"> 1. Light Post-editing: This level aims to make the translated content comprehensible and ensure that it accurately conveys the meaning of the original text. The focus is on correctness rather than style or fluency. 2. Full Post-editing: This level not only aims for an accurate translation but also improves the text's style, fluency, and coherence to achieve a quality similar to human translation. Full post-editing often requires the translator to revise the text so that it doesn't seem like it was machine-translated. <p>The type of post-editing chosen usually depends on the purpose of the translation and the quality standards required for the final text.</p> <p>The growing use of Machine Translation has made post-editing an increasingly important skill for professional translators. It allows for a more efficient translation process, especially for larger volumes of text and when a rough translation is sufficient. However, for texts requiring high quality, nuanced understanding, and cultural sensitivity, human translation is still necessary.</p>

Table A-9. CoTPs in task Language checking with ChatGPT 3 and ChatGPT 4

1 st Prompt	Identify terminology errors without providing a glossary.
ChatGPT 3 – 1 st Answer	<i>It failed – it asserts that terminology is correct even when some terms are omitted.</i>
ChatGPT 4 – 1 st Answer	In the Spanish text, the last sentence is translated as "En Estados Unidos, los máximos históricos diarios de nieve superan en dos a uno a los mínimos históricos." However, this is incorrect because it seems to be talking about snow records, when the original English text is referring to temperature records. A better translation would be "En los Estados Unidos, los récords diarios de altas temperaturas superan en dos a uno a los récords de bajas temperaturas." This now correctly conveys that record high temperatures are outnumbering record lows two-to-one, not record snowfalls.

2nd
Prompt Ensure the glossary has been considered when translating the text. If not, indicate when it has not been respected.

EN	ES
computer model	modelo informático
surface temperature	temperatura en superficie
ice sheet	manto de hielo
greenhouse gas	gas de efecto invernadero
fossil fuel	combustible fósil
sea ice	hielo marino
land area	superficie terrestre
ice core	núcleo helado
factory	factoría
sea surface	superficie marina

English: “Climate change is often cast as a prediction made by complicated computer models. But the scientific basis for climate change is much broader, and models are actually only one part of it (and, for what it’s worth, they’re surprisingly accurate).

For more than a century, scientists have understood the basic physics behind why greenhouse gases like carbon dioxide cause warming. These gases make up just a small fraction of the atmosphere but exert outsized control on Earth’s climate by trapping some of the planet’s heat before it escapes into space. This greenhouse effect is important: It’s why a planet so far from the sun has liquid water and life!

However, during the Industrial Revolution, people started burning coal and other fossil fuels to power factories, smelters and steam engines, which added more greenhouse gases to the atmosphere. Ever since, human activities have been heating the planet.

We know this is true thanks to an overwhelming body of evidence that begins with temperature measurements taken at weather stations and on ships starting in the mid-1800s. Later, scientists began tracking surface temperatures with satellites and looking for clues about climate change in geologic records. Together, these data all tell the same story: Earth is getting hotter.

Average global temperatures have increased by 2.2 degrees Fahrenheit, or 1.2 degrees Celsius, since 1880, with the greatest changes

happening in the late 20th century. Land areas have warmed more than the sea surface and the Arctic has warmed the most — by more than 4 degrees Fahrenheit just since the 1960s. Temperature extremes have also shifted. In the United States, daily record high snow outnumbers record lows two-to-one.”

Spanish: “El cambio climático a menudo se presenta como una predicción hecha por modelos de computadora complicados.

Pero la base científica del cambio climático es mucho más amplia, y los modelos son solo una parte de ella (y, por lo que vale, son sorprendentemente precisos).

Durante más de un siglo, los científicos han comprendido la física básica detrás de por qué los gases de efecto invernadero como el dióxido de carbono causan el calentamiento.

Estos gases representan solo una pequeña fracción de la atmósfera pero ejercen un control desproporcionado sobre el clima de la Tierra al atrapar parte del calor del planeta antes de que escape al espacio.

Este efecto invernadero es importante: ¡Es por eso que un planeta tan lejano del sol tiene agua líquida y vida!

Sin embargo, durante la Revolución Industrial, las personas comenzaron a quemar carbón y otros combustibles fósiles para alimentar fábricas, fundiciones y motores de vapor, lo que agregó más gases de efecto invernadero a la atmósfera.

Desde entonces, las actividades humanas han estado calentando el planeta.

Sabemos que esto es cierto gracias a un abrumador cuerpo de evidencia que comienza con las mediciones de temperatura tomadas en estaciones meteorológicas y barcos a partir de mediados del siglo XIX.

Más tarde, los científicos comenzaron a rastrear las temperaturas superficiales con satélites y buscar pistas sobre el cambio climático en registros geológicos.

Juntos, estos datos cuentan la misma historia: la Tierra se está calentando.

Las temperaturas globales promedio han aumentado en 2.2 grados Fahrenheit, o 1.2 grados Celsius, desde 1880, con los mayores cambios ocurriendo a finales del siglo XX. Las zonas terrestres se han calentado más que la superficie del mar, y el Ártico es la que más se ha calentado: más de 4 grados Fahrenheit sólo desde la década de 1960. Las temperaturas extremas también han cambiado. En Estados Unidos, los máximos históricos diarios de nieve superan en dos a uno a los mínimos históricos.”

ChatGPT 3 – 2 nd Answer	The glossary has been fully respected in the translation of the text.
ChatGPT 4 – 2 nd Answer	<p>The Spanish translation respects the provided glossary for most of the terms. However, there are two instances where it does not fully align with the glossary terms:</p> <ol style="list-style-type: none"> 1. "Computer models" has been translated as "modelos de computadora" instead of the glossary term "modelo informático". 2. "Surface temperatures" is translated as "temperaturas superficiales" instead of the glossary term "temperatura en superficie". 3. "Factories" is translated as "fábricas" instead of the glossary term "factoría". 4. "Sea surface" is translated as "la superficie del mar", instead of the glossary term "superficie marina". <p>These instances should be corrected to align with the glossary.</p>
3 rd Prompt:	Convert the measurement units to a different metric system.
ChatGPT 3 – 3 rd Answer	It failed – it only changes "Fahrenheit" to "Celsius," without recalculating the degrees.
ChatGPT 4 – 3 rd Answer	<p>Sure, I can convert the Fahrenheit degrees to Celsius degrees (which is used in the metric system), and the "quarter-mile" distance to kilometers:</p> <p>English: "Average global temperatures have increased by 2.2 degrees Fahrenheit, or 1.2 degrees Celsius, since 1880, with the greatest changes happening in the late 20th century. Land areas have warmed more than the sea surface and the Arctic has warmed the most — by more than 4 degrees Fahrenheit just since the 1960s. Temperature extremes have also shifted. In the United States, daily record high snow outnumbers record lows two-to-one."</p> <p>Converted English: "Average global temperatures have increased by 1.2 degrees Celsius, since 1880, with the greatest changes happening in the late 20th century. Land areas have warmed more than the sea surface and the Arctic has warmed the most — by more than 2.2 degrees Celsius just since the 1960s. Temperature extremes have also shifted. In the United States, daily record high snow outnumbers record lows two-to-one."</p> <p>Spanish: "Lo mismo ocurre con la casa de los Travers, cuando la encuentra, un cuarto de milla más allá."</p> <p>Converted Spanish: "Lo mismo ocurre con la casa de los Travers, cuando la encuentra, unos 0.4 kilómetros más allá."</p>

Table A-10. CoTPs in task Revision and Review with ChatGPT 3 and ChatGPT 4

1 st Prompt	Define translation review and revision.
ChatGPT 3 – 1 st Answer	<i>Accurately completed – it provides a full definition that includes both translation revision and target text review.</i>
ChatGPT 4 – 1 st Answer	<i>Accurately completed – it provides a full definition that includes both translation revision and target text review.</i>
2 nd Prompt	<p>Confirm that the translated text (English) is adequate with respect to the original (Spanish). These are the texts:</p> <p>Spanish: "LA COMISIÓN EUROPEA, Visto el Tratado de Funcionamiento de la Unión Europea, Visto el Reglamento (UE, Euratom) 2021/770 del Consejo, de 30 de abril de 2021, sobre el cálculo del recurso propio basado en los residuos de envases de plástico que no se reciclan, sobre los métodos y el procedimiento de puesta a disposición de dicho recurso propio, sobre las medidas para hacer frente a las necesidades de tesorería y sobre determinados aspectos del recurso propio basado en la renta nacional bruta (1), y en particular su artículo 5, apartado 6, Previa consulta al Comité creado con arreglo al artículo 39 de la Directiva 2008/98/CE del Parlamento Europeo y del Consejo (2), Considerando lo siguiente: (1) De conformidad con el artículo 5, apartado 5, del Reglamento (UE, Euratom) 2021/770, los Estados miembros deben enviar a la Comisión datos estadísticos sobre el peso en kilogramos de los residuos de envases de plástico generados y reciclados, así como el cálculo del importe del recurso propio basado en los residuos de envases de plástico que no se reciclan. (2) A fin de limitar la carga administrativa, los Estados miembros deben poder transmitir datos estadísticos y el importe del recurso propio en un único estado. (3) Los datos sobre la generación y el reciclado de residuos de envases de plástico constituyen la base para calcular las contribuciones nacionales al presupuesto general de la Unión. Por ello, es necesario que se refuerce la comparabilidad, la fiabilidad y la exhaustividad de esos datos. (4) A fin de garantizar la comparabilidad, fiabilidad y exhaustividad de los datos compartidos entre los Estados miembros, conviene establecer normas detalladas sobre los datos que han de figurar en el estado que debe facilitarse a la Comisión."</p> <p>English: "THE EUROPEAN COMMISSION, Having regard to the Treaty on the Functioning of the European Union, Having regard to</p>

	<p>Council Regulation (EU, Euratom) 2021/000 of 30 April 2021 on the calculation of the own resource based on plastic packaging waste that is not recycled, on the methods and procedure for making available that own resource, on the measures to meet cash requirements, and on certain aspects of the own resource based on gross national income (1), and in particular Article 5(0) thereof, After consulting the committee established by Article 00 of Directive 2008/98/EC of the European Parliament and of the Council (2), Whereas: (1) Pursuant to Article 5(0) of Regulation (EU, Euratom) 2021/000, Member States should send to the Commission statistical data on the weight in kilograms of generated and recycled plastic packaging waste, as well as the calculation of the amount of the own resource based on non-recycled plastic packaging waste. (2) In order to limit the administrative burden, Member States should be able to transmit statistical data and the amount of the own resource in a single statement. (3) Data on plastic packaging waste generation and recycling constitute the basis for calculating the national contributions to the general budget of the Union. Therefore, it is necessary to reinforce the comparability, reliability and exhaustiveness of those data. (4) In order to guarantee the comparability, reliability and exhaustiveness of data between the Member States, it is appropriate to lay down detailed rules concerning the data to be contained in the statement to be provided to the Commission."</p>
<p>ChatGPT 3 – 2nd Answer</p>	<p>Yes, the translated English text is adequate with respect to the Spanish original. The English text accurately reflects the content of the Spanish text, including technical terms and legal provisions. The structure and tone of the text is also maintained in the translation.</p>
<p>ChatGPT 4 – 2nd Answer</p>	<p>Based on the provided Spanish and English texts, the English translation seems to accurately reflect the content of the original Spanish version. Key elements like the reference to the Treaty on the Functioning of the European Union, Council Regulation, the consultation of the committee established by a directive, and the articles concerning plastic waste and national contributions are all well accounted for.</p> <p>However, a slight discrepancy can be noticed. In the Spanish text, the phrase "Reglamento (UE, Euratom) 2021/770" is mentioned, while in the English translation, the same has been mentioned as "Regulation (EU, Euratom) 2021/000." If "2021/770" is the correct reference number for the Regulation, the English version should reflect this as well. This number should be checked for accuracy. If this was an error in the English text, the English version should read "Regulation (EU, Euratom) 2021/770" for full consistency with the Spanish text.</p> <p>Except for this minor discrepancy, the overall translation from Spanish to English appears to be done adequately, as it seems to preserve the intent, tone, and meaning of the original text.</p>

Table A-11. CoTPs in task Regular expression with ChatGPT 3 and ChatGPT 4.

1 st Prompt	Give me information on the abusive use of adverbs ending in <i>-mente</i> .
ChatGPT 3 – 1 st Answer	<p>The abusive use of adverbs ending in "-mente" is a common problem in writing, especially for beginning writers. These adverbs are used to modify verbs, adjectives, or other adverbs and, although they are useful for accurately describing actions and states, their overuse can make the text redundant, repetitive and uninteresting for the reader.</p> <p>Excessive use of adverbs ending in "-mente" can also affect the clarity and flow of the text, as they can disrupt the rhythm of the reading and make the prose seem stilted and unnatural. It is important to remember that there are other ways of describing actions and states, such as the use of adjectives, nouns, and more precise verbs.</p>
ChatGPT 4 – 1 st Answer	<p>Adverbs ending in "-mente" in Spanish, similar to "-ly" in English, are frequently used to modify verbs, adjectives, and other adverbs. They can provide useful nuances in the sentence; however, their excessive use can be considered a stylistic problem in writing, particularly in formal or academic contexts.</p> <p>Here's why the abusive use of "-mente" adverbs can be problematic:</p> <ol style="list-style-type: none"> 1. Redundancy: Many times, the adjective itself carries the full meaning and the adverb adds nothing but redundancy. For example, instead of saying "absolutamente seguro" (absolutely sure), you can simply say "seguro" (sure). 2. Monotonous and cumbersome reading: When overused, these adverbs can make the text monotonous and more difficult to read. Variety in sentence structure and word choice tends to make prose more engaging and dynamic. 3. Distracts from the main point: Over-reliance on "-mente" adverbs can distract from the main point of the sentence. The main focus should be on strong verbs and nouns, not on the adverbs modifying them. <p>To avoid overuse of "-mente" adverbs:</p> <ul style="list-style-type: none"> - Use synonyms: There are many synonyms that can replace adverbs and make the text more varied and interesting. - Remove unnecessary adverbs: If the meaning of the sentence remains clear without the adverb, you can simply eliminate it. - Use other sentence structures: Instead of relying on "-mente" adverbs, you can create sentences using prepositions, conjunctions, and other structures to express the same idea. <p>Remember, these are stylistic suggestions. The use of "-mente" adverbs isn't wrong per se, but like any other linguistic resource, they should be used judiciously.</p>

2 nd Prompt	I have a text in which the use of adverbs ending in -mente is disproportionate. Edit it to make its use appropriate: <i>Ligeramente volteado a estribor, de su arboladura intacta colgaban tranquilamente las piltrafas escuálidas del velamen, entre jarcias adornadas elegantemente de orquídeas. El casco, cubierto con una tersa coraza de rémora petrificada y musgo tierno, estaba firmemente enclavado en un suelo de piedras</i>
ChatGPT 3 – 2 nd Answer	Sutilmente volteado a estribor, de su arboladura intacta colgaban serenamente las escasas telas desgarradas, mientras las jarcias se engalanaban elegantemente con orquídeas. El casco, protegido por una suave coraza de rémora petrificada y musgo delicado, se encontraba firmemente encajado en un lecho de piedras.
ChatGPT 4 – 2 nd Answer	Your text, edited to reduce the use of "-mente" adverbs, could be written as follows: Con una ligera inclinación a estribor, las escuálidas piltrafas del velamen colgaban sin alteración de su arboladura intacta, entre jarcias adornadas con orquídeas de forma elegante. El casco, cubierto por una coraza tersa de rémora petrificada y musgo suave, se encontraba anclado firmemente en un suelo de piedras.
	Here's how I made the changes: - Changed "ligeramente volteado" to "con una ligera inclinación". The adverb "ligeramente" is eliminated by incorporating the sense of slightness into the noun phrase. - Changed "tranquilamente las piltrafas escuálidas del velamen" to "las escuálidas piltrafas del velamen colgaban sin alteración". By using a phrase ("sin alteración") that conveys the tranquility without needing to use an adverb. - Changed "adornadas elegantemente de orquídeas" to "adornadas con orquídeas de forma elegante". Instead of the adverb modifying the verb directly, I've used a prepositional phrase to convey the manner of the decoration. - Left "firmemente enclavado" as it is, because this seems like a crucial part of the description and changing it might affect the intended meaning. Remember that these are just suggestions and the optimal changes might depend on the specific stylistic and tonal requirements of your text.
3 rd Prompt	Provide it as Regex syntax for use in the memoQ program.
ChatGPT 3 – 3 rd Answer	Here I provide you with the regular expression to search for words ending in "-mente" in the text: <code>\b\w+mente\b</code> And here I provide you with the regular expression to search for and replace all words ending in "-mente" in the text: Search: <code>\b(\w+)mente\b</code>

	<p>Replace with: \$1</p> <p>This regular expression will search for all words ending in "-mente" and replace them with the same word without the "-mente" suffix</p>
ChatGPT 4 – 3 rd	To apply regular expressions (regex) for matching adverbs ending in "-mente" in Spanish within memoQ, you could use a regex expression like the following:
Answer	<pre> ... \b\w+mente\b ... </pre> <p>Here's what each element of the expression does:</p> <ul style="list-style-type: none"> - `b` is a word boundary. It makes sure we're only getting whole words, not words that contain the pattern inside them. - `w+` matches one or more word characters (that's the `w`), where "one or more" is specified by the `+`. - `mente` matches the actual string "mente", which is the common ending of adverbs in Spanish. - `b` at the end is another word boundary, again ensuring we're only getting whole words. <p>This regex should match any whole word that ends in "mente".</p> <p>Note: Regular expressions can be very powerful, but also tricky. Always test your expressions before applying them to critical work, to make sure they're doing what you want.</p> <p>Remember to use the Regex Tagger in memoQ to tag the matched segments. You can use this as a quality assurance step to find all instances of "-mente" adverbs in your translations.</p>

Language Complexity in Human and Machine Translation: A Preliminary Study

Gábor Recski¹[0000–0001–5551–3100] and Fanni Kádár²

¹ TU Wien, Vienna, Austria

gabor.recski@tuwien.ac.at

² Deloitte Hungary

fkadar@deloittece.com

Abstract. Systematic comparison between machine translation (MT) and human translation (HT) is mostly limited to the evaluation of MT output with HT as reference, as opposed to a more general study of the properties of MT and HT output texts. We present preliminary experiments investigating differences between MT and HT in terms of readability and language complexity. We perform both quantitative and qualitative comparison of the outputs of machine and human translation, using samples of English text across multiple domains and genres and their Hungarian translations created by humans and by the state-of-the-art machine translation system `deepl`. Our results show that machine translation produces somewhat simpler text than human translation on 3 out of 4 samples, and on 2 samples this effect is caused primarily by human translators using a higher number of complex words. We release all software used in our experiments to facilitate further studies on larger samples, additional languages and domains, and using alternative MT systems.

Keywords: Machine Translation, Language Complexity, Readability

1 Introduction

Recent advances in machine translation (MT) have fueled the debate about the future role of human translation (HT), yet systematic comparison between MT and HT is still mostly limited to the evaluation of MT output with HT as reference, as opposed to a more general study of the properties of MT and HT output texts. With this preliminary work we propose to investigate differences between MT and HT in terms of the closely related concepts *readability* and *language complexity*. We perform both quantitative and qualitative comparison of the outputs of machine and human translation, using samples of English text across multiple domains and genres, each of which have been translated into Hungarian both by a human and by the state-of-the-art MT system `deepl`. Our results suggest that machine translation produces somewhat simpler text than human translation on 3 out of 4 samples, and that this effect is often caused by human translators using a higher number of complex words.

The remainder of this paper is structured as follows. Section 2 summarizes related work, Section 3 describes our data sources and documents the process of creating the samples. Section 4 describes the metrics used to analyze each sample, Section 5 presents and discusses the results, and Section 6 concludes the paper. All data and software described in this paper is publicly available under an MIT license from GitHub³.

2 Related Work

Most recent work on the evaluation of machine translation (MT), whether it is based on comparison with human reference translations or on direct assessment by human evaluators, is focused on ranking MT systems from best to worst, as dictated by the leaderboard culture dominating the field of natural language processing (NLP) [14,6]. In this context, comparison of MT output to human translation (HT) is usually performed only quantitatively and with the sole purpose of measuring the degree of similarity between MT and HT output. Exceptions include the work of Ahrenberg [1], who performs a close qualitative comparison of MT and HT on a single newspaper article translated from English into Swedish, focusing on information structure, word order, post-editing required for the MT output, and examples of Translation Procedures [16]. The simple quantitative analysis found that the MT output more closely preserved the number of sentences and tokens of the English original than the human translator, who would often split longer sentences into multiple shorter ones. Another direct comparison of MT and HT is a detailed qualitative comparison of French subtitles of a television show translated from English to French both by `deep1` and by professional translators [2], the quantitative component of this analysis focused on the categorization of errors in the MT output.

Another recent analysis compares HT with the output of post-editing (PE) of MT output [23], this work involves statistical analysis of samples across 5 languages and multiple domains, focusing on part-of-speech (POS) sequences, text length, lexical diversity, and lexical density, the latter defined as the ratio of content words. These experiments demonstrate that PE output exhibits lower lexical variety and lexical density than MT, that average sentence length in PE output is closer to that in the source text than to that in the HT output, and that POS sequences in PE output are more similar to those typical of the source language. Another quantitative comparison of MT and HT with respect to lexical diversity was performed in [26], using the Europarl corpus [15] and showing for two language pairs and a wide variety of MT systems that MT produces text with lower lexical richness than HT. [25] extends this study to include measures of morphological variation, focusing on inflectional paradigms of lemmas and showing that MT reduces their diversity compared to HT across all language pairs.

The evaluation of text readability has a long tradition in both applied linguistics and NLP [13,11] Commonly used quantitative measures of readability such

³ <https://github.com/recki/comp-trans>

as the Flesch-Kincaid test [5,12] and Gunning’s Fog Index [8,7] rely on simple statistics about the average length of words and sentences, such common measures differ only in the exact methods for quantifying these two properties and in the parameters for combining them to obtain a single dimension of readability. Studies on language complexity also use measures such as type-token ratio and word entropy [27], but their relevance to readability remains unclear.

3 Data

We extracted 4 samples of English text from various domains and genres, each taken from publicly available sources. The TED3 sample is the text of three short talks given at various TED conferences, transcripts of which are available online⁴ with translations in multiple languages, including Hungarian. For creating the sample we used a preprocessed dataset⁵ that has been previously used for improving neural machine translation systems [21]. Our sample contains the complete transcript of two presentations. The first was given by Barat Ali Batoor, the Hungarian translation was created by Zsuzsanna Dan and reviewed by Orsi Mance. The second talk is by Kees Moliker, the Hungarian translator of the transcript was Mária Ruzsáné Cseresnyés, the reviewer of the translation was Anna Patai.

Our second sample, FGM, is an excerpt from the script of the movie *A few good men*, both English and Hungarian texts were extracted from SRT-formatted subtitles downloaded from OpenSubtitles⁶. The English text is an exact transcript of the movie dialogue, the Hungarian subtitles were extracted from the DVD edition, the translator could not be identified. The data was converted to raw text using the `srt` library⁷. The 1984 sample is an excerpt from the novel *1984* by George Orwell. The original text was downloaded from Project Gutenberg⁸, the Hungarian version (translated by László Szíjgyártó) was extracted from the Szeged Treebank [3]. DC567 is a single document from the multilingual JRC-ACQUIS corpus [22], a 2006 communication of the European Commission titled *A Contribution to the EU’s Growth and Jobs Strategy*.

Basic statistics describing the original English and the two Hungarian translations of each of the four samples is presented in Table 1. The four samples were selected to be comparable in length, as measured by the number of words in the English original text. Differences in average sentence length (number of words per sentence) and relative vocabulary size (number of unique lemmas compared to the number of words, also known as the type-token ratio, as measured by Herdan’s C, see Section 4 for details) are characteristic of each genre. When comparing the English originals, the novel *1984* contains the longest sentences and has the largest vocabulary. The *DC567* text, a policy statement on matters

⁴ <https://www.ted.com/participate/translate>

⁵ <https://github.com/neulab/word-embeddings-for-nmt>

⁶ <https://www.opensubtitles.org/>

⁷ <https://pypi.org/project/srt/>

⁸ <http://gutenberg.net.au/ebooks01/0100021.txt>

of the economy, is a close second in average sentence length but has a smaller vocabulary, comparable to the presentations in *TED3*. The movie dialogues in *FGM* have the shortest sentences and the smallest vocabulary.

Table 1. Basic statistics of the four samples. *sens* and *words* are the total numbers of sentences and words in each sample, respectively. *w/s* is the average number of words per sentence (including punctuation), $|V|$ is the vocabulary size measured as the number of unique lemmas, C_l is Herdan’s C on the set of lemmas (see Section 4 for details).

		sens	words	w/s	$ V $	C_l
1984	EN	258	6181	23.96	1428	.83
	HU (human)	251	5040	20.08	1641	.87
	HU (deepl)	257	5077	19.75	1630	.87
TED3	EN	320	4940	15.52	990	.81
	HU (human)	298	4086	13.71	1223	.85
	HU (deepl)	303	4247	14.02	1178	.85
FGM	EN	753	7241	9.62	1106	.79
	HU (human)	562	3680	6.55	1095	.85
	HU (deepl)	1056	6244	5.91	1444	.83
DC567	EN	277	6175	22.29	1059	.80
	HU (human)	260	5845	22.48	1460	.84
	HU (deepl)	262	5798	22.13	1338	.83

In addition to the existing human translations we translated all four samples from English into Hungarian using `deepl`. Each sample was passed to `deepl` in a single API⁹ call, with the entire text as a single `text` parameter, using default values for all other parameters. The API was accessed using the `deepl-python` client¹⁰. Each MT output as well as the human translations and the English originals were tokenized and lemmatized using the default models of `stanza`¹¹ [20], accessed via a wrapper¹² provided by the `tuw-nlp` library.

4 Methods

We compute for each of our samples simple statistical measures of language complexity. A common measure of lexical diversity is the type-to-token ratio (TTR) of a document, measuring the ratio between the number of distinct lemmas (types) and the total number of words in the text (tokens). A common version

⁹ <https://www.deepl.com/docs-api/translate-text/markup/>

¹⁰ <https://github.com/DeepLcom/deepl-python>

¹¹ <https://stanfordnlp.github.io/stanza/>

¹² https://github.com/recski/tuw-nlp/blob/main/scripts/run_pipeline.py

of this measure is Herdan’s C [24], also used to measure language complexity by [27], calculated as

$$\frac{\log \#types}{\log \#tokens}$$

For each document we calculate Herdan’s C both on the set of words and on the set of lemmas. Since the type-token ratio is highly sensitive to text length, we also compute the Measure of Textual Lexical Diversity (MTLD) [17,18] for each sample, which measures the mean length of word sequences that maintain a given TTR value (0.72 by default). We calculate MTLD on both tokens and lemmas, using the implementation available from the `lexical-diversity` python package¹³.

Finally, we also calculate Gunning’s Fog Index [8,7], a common measure of text readability based on average sentence length and the ratio of complex words, the latter defined as words containing 3 or more syllables. The formula for calculating the Fog Index is the following

$$0.4 \times \left(\frac{\#words}{\#sentences} + 100 \times \frac{\#complex\ words}{\#total\ words} \right)$$

While this metric is widely used for measuring the readability of English text, its applicability to Hungarian is yet to be investigated (see [4] for a first discussion). For this study we used two large text corpora of English and Hungarian to measure the distribution of word length (as measured by the number of syllables) and approximate a more realistic language-specific parameter of the FOG index for Hungarian. For English, we used word counts extracted from the UMBC-Webbase¹⁴ corpus [9]. Syllables of each word were counted by lemmatizing them using `stanza` (see Section 3) and retrieving the number of syllables for each word from the CMU pronunciation dictionary¹⁵ provided by the Python library `pronouncing`¹⁶. Since the dictionary’s coverage is limited, but the full vocabulary of the corpus contains 6.7 million word types, we only lemmatized words with at least 100 occurrences in the UMBC corpus (175 000 types) and then proceeded to further filter out those lemmas that are not present in the CMU dictionary. The remaining set of types still covers 2.89 billion of the 3.34 billion tokens in the UMBC corpus. For Hungarian we used a version of the Hungarian Webcorpus¹⁷ [19] that has been lemmatized using `emtsv`¹⁸ [10]. Since the number of syllables in a Hungarian word can be determined by simply counting the number of vowel characters, we could efficiently process all 8.8 billion tokens in the corpus.

The distribution of the syllable counts in the two corpora is presented in Figure 1, the raw data is available from the project repository. We observe that by considering all English words with 3 or more syllables as complex the Gunning

¹³ <https://pypi.org/project/lexical-diversity/>

¹⁴ <https://ebiquity.umbc.edu/resource/html/id/351/UMBC-webbase-corpus>

¹⁵ <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>

¹⁶ <https://pypi.org/project/pronouncing/>

¹⁷ <https://hlt.bme.hu/en/resources/webcorpus2>

¹⁸ <https://github.com/nytud/emtsv>

FOG index treats about 17% of all tokens in a large webcorpus as complex, while for Hungarian a similar ratio (about 15%) can be achieved by considering words with 4 or more syllables as complex. Therefore in the present study we shall calculate the Gunning FOG index for Hungarian texts using this adjusted parameter, noting that further quantitative and qualitative studies would be required to determine whether this metric can be considered a reliable measure of the readability of Hungarian text.

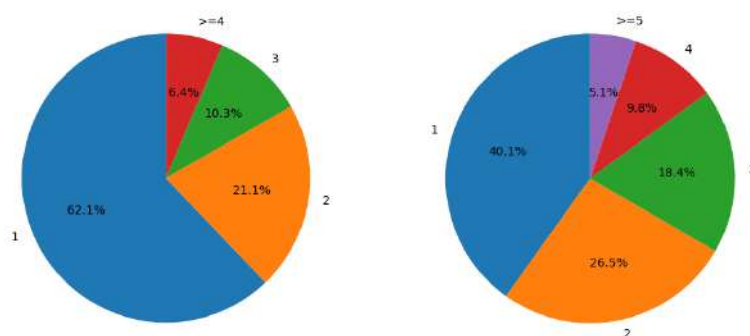


Fig. 1. Distribution of the number of syllables in words of the UMBC-Webbase corpus (left) and the Hungarian Webcorpus (right)

5 Results

The complexity measures described in Section 4 are shown in Table 2. We observe that while Herdan’s C values show little or no difference between human and machine translations of the same text, the MTDL measure is considerably lower for MT outputs across all samples, suggesting that MT produces lexically less diverse text. Additionally, the Gunning Fog Index is lower for texts produced by `deep1` than for those produced by human translators on 3 of the 4 samples, suggesting that machine translation generates somewhat simpler text across a variety of genres and domains. The two factors and how they contribute to Gunning’s F for each sample are visualized in Figure 2, showing that while on the 1984 and DC567 samples the difference is mainly due to the MT output using fewer complex words, on FGM it is primarily caused by shorter average sentences.

We also inspected each sample manually to understand possible sources of quantitative differences between each pair of human and machine translation. First we observe that in the FGM sample of movie subtitles the human translator considerably altered and shortened the contents of the original text, most likely to reduce the number of characters that viewers have to read over a given period

Table 2. Complexity measures for the four samples. w/s is the average number of words per sentence, L_c/L is the ratio of complex lemmas, C and M are Herdan's C and MTLT, each measured on both the set of all words (C_w , M_w) and all lemmas (C_l , M_l). F is Gunning's Fog Index. For details of each of these measures see Section 4

		w/s	L_c/L	C_w	C_l	M_w	M_l	F
1984	EN	23.96	6.76%	.85	.83	77.7	57.9	12.29
	HT	20.08	6.77%	.90	.87	105.6	72.6	10.74
	MT	19.75	6.36%	.89	.87	82.1	58.6	10.45
TED3	EN	15.52	5.98%	.83	.81	57.0	39.5	8.60
	HT	13.71	5.58%	.89	.85	67.1	46.5	7.72
	MT	14.02	6.19%	.89	.85	55.2	38.2	8.08
FGM	EN	9.62	4.17%	.82	.79	62.6	43.8	5.51
	HT	6.55	4.24%	.89	.85	54.4	40.0	4.31
	MT	5.91	3.91%	.87	.83	38.9	30.9	3.93
DC567	EN	22.29	18.38%	.83	.80	89.0	68.5	16.27
	HT	22.48	19.26%	.89	.84	109.7	71.4	16.70
	MT	22.13	18.32%	.88	.83	93.9	59.2	16.18

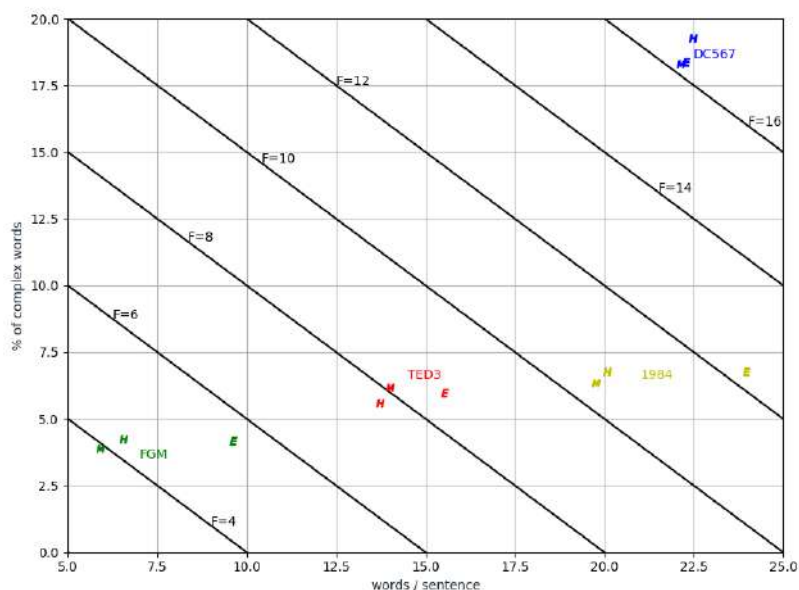


Fig. 2. Contributing factors of Gunning's Fog Index. On each sample, E , H , and M show values for the English original, the human translation, and the machine translation, respectively.

of time, a standard practice in the production of movie subtitles [2]. This practice is illustrated by the excerpt in Figure 3. As a result, the Hungarian subtitles contain only about half as many sentences and words as the machine translation of the English subtitles (see Table 1 in Section 3), calling into question whether the two texts should be considered comparable.

Next we observed that the differences in average sentence length in 1984 are primarily due to the fact that the human translator is more likely to merge multiple English sentences into a single Hungarian sentence. In the 1984 sample 258 English sentences are translated into 257 and 251 sentences by `deep1` and the human translator, respectively, a difference of 2.3%, while the difference in the number of tokens (5077 and 5040) is only 0.7%. A closer inspection of the samples revealed that most of this difference was caused by a single decision of the human translator to translate the five sentences in Figure 4 as a single sentence. Finally we looked at examples of MT using fewer complex words than the human translator. Recall that we consider a word complex if it has 4 or more syllables, based on the calculation presented in Section 4. Figure 5 shows an example sentence from the DC567 sample in English and in the two Hungarian translations, illustrating the human translators' preference for longer words.

6 Conclusion

We presented a preliminary study comparing human and machine translations of the same texts from the point of view of language complexity and readability. Samples of English text from four domains were extracted with existing Hungarian translations written by humans, the machine translation system `deep1` was used to create machine translations. HT and MT output was compared quantitatively using multiple measures of lexical diversity and Gunning's Fog Index of readability. For Hungarian we have estimated a language-specific parameter for calculating Gunning's F on the basis of the distribution of word length in large corpora of English and Hungarian. Qualitative analysis of the samples was also performed to uncover reasons for the observed quantitative differences. We find that on 3 out of 4 samples human translations exhibit slightly higher complexity levels than machine translations. On the two samples containing fiction and legal text this effect is the result of the human translator choosing more complex words, while in the case of the sample containing movie subtitles it appears to be a byproduct of the human translator shortening the text to about half its length, most likely in compliance with external requirements, a finding that questions the suitability of movie subtitles for this type of study.

In light of these preliminary results we believe that further study is required to determine whether MT and HT output exhibit systematic differences in readability and whether such effects are specific to certain domains, genres, or languages. When interpreting measures of readability, their applicability to a given language should also be investigated. We release all code used in the paper as

Original subtitles	Human translation
<i>Daniel Alistair Kaffee, born June 8, 1964... at Boston Mercy Hospital.</i>	<i>Daniel A. Kaffee, született 1964. június 8-án, Bostonban.</i> 'Daniel A. Kaffee, born June 8, 1964, in Boston.'
<i>Your father's Lionel Kaffee... former Navy judge advocate and attorney general of the United States. Died 1985.</i>	<i>Apja Lionel Kaffee, korábbi igazságügyi miniszter.</i> 'Your father was Lionel Kaffee, former Attorney General'
<i>You went to Harvard law. Then you joined the Navy... probably because that's what your father wanted you to do.</i>	<i>A Harvardra járt, és az apja kívánságára jött ide.</i> 'You went to Harvard, and came here at the request of your father'
<i>And now you're just treading water for three years in the JAG Corps. Just laying low till you can get out and get a real job.</i>	<i>Csak tölti itt az idejét, míg be nem jön a jól fizető állás.</i> 'You are just killing time until a well-paying job comes along'
<i>If that's the situation, that's fine. I won't tell anyone.</i>	<i>Ám legyen! Nem árulom el senkinek.</i> 'So be it! I won't tell anyone.'
<i>But it's my feeling that if this case is handled in the same... fast-food, slick-ass, Persian bazaar manner... with which you seem to handle everything else... then something's gonna get missed.</i>	<i>De ha a szokásos bazári módján kezeli ezt az ügyet...'</i> 'But if you're going to handle this case in your usual bazaar manner..' <i>...akkor épp a lényeg szikkad el.</i> '... you will miss the point'
<i>And I wouldn't be doing my job if I allowed Dawson and Downey... to spend any more time in prison than absolutely necessary... because their attorney had predetermined the path of least resistance.</i>	<i>De nem hagghatom, hogy a börtönben aszalják őket...'</i> 'But I cannot let them rot in prison..' <i>...mert a védőjük a legkisebb ellenállást kedveli.</i> '... just because their attorney prefers the least resistance'

Fig. 3. Excerpt from the FGM sample illustrating that the human translator considerably shortened the subtitles, most likely to comply with requirements on the maximum amount of text that can be displayed over a certain period of time. In this sample, 185 tokens in the original text were reduced to just 91 tokens in the translation.

They were the homes of the four Ministries between which the entire apparatus of government was divided. The Ministry of Truth, which concerned itself with news, entertainment, education, and the fine arts. The Ministry of Peace, which concerned itself with war. The Ministry of Love, which maintained law and order. And the Ministry of Plenty, which was responsible for economic affairs.

Fig. 4. A series of five sentences from the 1984 sample that was mapped to a single Hungarian sentence by the human translator

open-source software¹⁹ to facilitate similar studies on additional languages and domains, on larger samples, or using alternative MT systems.

Acknowledgements We thank the two anonymous reviewers for their many useful suggestions, for pointing out additional references, and for recommending MTDL as an additional measure of lexical diversity.

¹⁹ <https://github.com/recski/comp-trans>

EN: *The Doha Development Agenda remains our first priority*

HT: *A Dohai Fejlesztési Programterv marad elsődleges prioritásunk,*
the Doha-ATTR development-ATTR agenda remain-3SG primary priority-2PL.POSS

MT: *A dohai fejlesztési menetrend továbbra is első számú prioritásunk,*
the Doha-ATTR development-ATTR agenda further too first number-ED priority-2PL.POSS

EN: *and the Commission is working intensively to restart the Doha negotiation*

HT: *ezért a Bizottság nagy erőfeszítéseket tesz a dohai tárgyalások újraindítása érdekében,*
therefore the Commission great effort-PL-ACC make-3SG the Doha-ATTR negotiation-PL restart-3SG.POSS.interest-3SG.POSS-INE

MT: *és a Bizottság intenzíven dolgozik a dohai tárgyalások újraindításán*
and the Commission intensive-ADV work-3SG the Doha-ATTR negotiation-PL restart-3SG.POSS-SUP

EN: *after its suspension in July 2006.*

HT: *azok 2006. júliusi feljegyzését követően.*
those 2006. July-ATTR suspension-3SG.POSS-ACC follow-PART-ADV

MT: *miután 2006. júliusában felfüggesztették azokat.*
after 2006. July-3SG.POSS-INE suspend-3PL:DEF those-ACC

Fig. 5. Sample sentence from the DC567 data, illustrating the human translator's preference for longer words. Bold indicates Hungarian words with 4 or more syllables in their lemma, considered as complex when calculating Gunning's F (see Section 4)

References

1. Ahrenberg, L.: Comparing machine translation and human translation: A case study. In: Proceedings of the Workshop Human-Informed Translation and Interpreting Technology. pp. 21–28. Association for Computational Linguistics, Shoumen, Bulgaria, Varna, Bulgaria (2017). https://doi.org/10.26615/978-954-452-042-7_003
2. Baldo de Brébisson, S.: Comparison between automatic and human subtitling: A case study with game of thrones. In: Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019). pp. 1–10. Incoma Ltd., Shoumen, Bulgaria, Varna, Bulgaria (2019). https://doi.org/10.26615/issn.2683-0078.2019_001
3. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Lecture Notes in Computer Science: Text, Speech and Dialogue. pp. 123–131. Springer (2005)
4. Dóra, L.: Olvashatósági tesztek: elmélet és gyakorlat [readability tests: theory and practice]. KÉPZÉS ÉS GYAKORLAT: TRAINING AND PRACTICE **17**(2), 21–34 (2019)
5. Flesch, R.: How to write plain English. Harper and Row (1979)
6. Freitag, M., Rei, R., Mathur, N., Lo, C.k., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., Martins, A.F.T.: Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In: Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 46–68. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (2022), <https://aclanthology.org/2022.wmt-1.2>
7. Gunning, R.: The Fog Index after twenty years. *Journal of Business Communication* **6**(2), 3–13 (1969)
8. Gunning, R., et al.: Technique of clear writing (1952)
9. Han, L., L. Kashyap, A., Finin, T., Mayfield, J., Weese, J.: UMBC_EBIQUITY-CORE: Semantic textual similarity systems. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity. pp. 44–52. Association for Computational Linguistics, Atlanta, Georgia, USA (2013), <https://aclanthology.org/S13-1005>
10. Indig, B., Sass, B., Simon, E., Mittelholcz, I., Vadász, N., Makrai, M.: One format to rule them all – the emtsv pipeline for Hungarian. In: Proceedings of the 13th Linguistic Annotation Workshop. pp. 155–165. Association for Computational Linguistics, Florence, Italy (2019). <https://doi.org/10.18653/v1/W19-4018>
11. Kanungo, T., Orr, D.: Predicting the readability of short web summaries. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining. p. 202–211. WSDM '09, Association for Computing Machinery, New York, NY, USA (2009). <https://doi.org/10.1145/1498759.1498827>
12. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep., Naval Technical Training Command Millington TN Research Branch (1975)
13. Klare, G.R.: Assessing readability. *Reading research quarterly* pp. 62–102 (1974)
14. Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., Graham, Y., Grundkiewicz, R., Haddow, B., Knowles, R., Koehn, P., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Novák, M., Popel, M.,

- Popović, M.: Findings of the 2022 conference on machine translation (WMT22). In: Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 1–45. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (2022), <https://aclanthology.org/2022.wmt-1.1>
15. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of Machine Translation Summit X: Papers. pp. 79–86. Phuket, Thailand (2005), <https://aclanthology.org/2005.mtsummit-papers.11>
 16. van Leuven-Zwart, K.: Translation and original: Similarities and dissimilarities, i. Target. *International Journal of Translation Studies* **1**(2), 151–181 (1989). <https://doi.org/https://doi.org/10.1075/target.1.2.03leu>
 17. McCarthy, P.M.: An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). Ph.D. thesis, The University of Memphis (2005)
 18. McCarthy, P.M., Jarvis, S.: MtlD, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* **42**(2), 381–392 (May 2010). <https://doi.org/10.3758/BRM.42.2.381>
 19. Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D. thesis, Eötvös Loránd University (2020), https://hlt.bme.hu/media/pdf/nemeskey_thesis.pdf
 20. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 101–108. Association for Computational Linguistics, Online (2020). <https://doi.org/10.18653/v1/2020.acl-demos.14>
 21. Qi, Y., Sachan, D., Felix, M., Padmanabhan, S., Neubig, G.: When and why are pre-trained word embeddings useful for neural machine translation? In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 529–535. Association for Computational Linguistics, New Orleans, Louisiana (2018). <https://doi.org/10.18653/v1/N18-2084>
 22. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06). European Language Resources Association (ELRA) (2006), http://www.lrec-conf.org/proceedings/lrec2006/pdf/340_pdf.pdf
 23. Toral, A.: Post-edited: an exacerbated translationese. In: Proceedings of Machine Translation Summit XVII: Research Track. pp. 273–281. European Association for Machine Translation, Dublin, Ireland (2019), <https://aclanthology.org/W19-6627>
 24. Tweedie, F.J., Baayen, R.H.: How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities* **32**, 323–352 (1998)
 25. Vanmassenhove, E., Shterionov, D., Gwilliam, M.: 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. pp. 2203–2213. Association for Computational Linguistics, Online (2021). <https://doi.org/10.18653/v1/2021.eacl-main.188>
 26. Vanmassenhove, E., Shterionov, D., Way, A.: Lost in translation: Loss and decay of linguistic richness in machine translation. In: Proceedings of Machine Translation Summit XVII: Research Track. pp. 222–232. European Association for Machine Translation, Dublin, Ireland (Aug 2019), <https://aclanthology.org/W19-6622>

27. Yasseri, T., Kornai, A., Kertész, J.: A practical approach to language complexity: a Wikipedia case study. PLoS ONE 7(11) (2012). <https://doi.org/10.1371/journal.pone.0048386>

Human Evaluation for Translation Quality of ChatGPT: A Preliminary Study

Yanqing Zhao, Min Zhang, Xiaoyu Chen, Yadong Deng, Aiju Geng, Limin Liu, Xiaoqin Liu, Wei Li, Yanfei Jiang, Hao Yang, Yu Han, Shimin Tao, Ning Xie, Xiaochun Li, Miaomiao Ma, and Zhaodi Zhang

Huawei Translation Services Center, Beijing, China

{zhaoyanqing, zhangmin186, chenxiaoyu35, dengyadong, gengaiju, cecilia.liulimin, liuxiaoqin09, vicky.li, jiangyanfei, yanghao30, hanyu630, taoshimin, nicolas.xie, carol.lixiaochun, mamiaomiao, zhangzhaodi}@huawei.com

Abstract. ChatGPT has shown promising results for Machine Translation (MT). However, whether it is comparable to standard translation models and performs well in some specific domain remains as an open question. In this paper, we conduct human evaluations on its translation performance in three domains using the Direct Assessment (DA) method. The evaluation result shows that ChatGPT as a whole achieves comparable performance with standard translation models, especially in the general domain. However, ChatGPT's performance is inferior in terms of translating domain-specific terminologies, and it appears to be informal when it comes to the Information and Communications Technology (ICT) and biomedical domains, where the text style is distinctively different from that in the general domain.

Keywords: ChatGPT, Human Evaluation, Direct Assessment.

1 Introduction

Recently, the emergence of ChatGPT has brought remarkable influence on Natural Language Processing (NLP) tasks. ChatGPT is an intelligent chatbot developed by OpenAI. It has shown promising results for Machine Translation (MT) and is becoming a new paradigm for translation [3, 2, 7].

To the best of our knowledge, the evaluations of ChatGPT translation quality in these studies [3, 2, 7] are all made based on commonly used automatic evaluation metrics (BLEU [6], COMET [8]), rather than human evaluation. We are particularly interested in human evaluation for ChatGPT translation results, especially the human evaluation gap with commercial translation systems in different domains.

In this paper, we conduct human Direct Assessment (DA) [1] on ChatGPT's translations and those of two other translation systems (Google Translate and Huawei Translate) in three domains, including one general domain (WMT22 General MT Task) and two specific domains (WMT22 biomedical and Information and Communications Technology (ICT) domains). The DA scoring criteria in

Table 1. DA scoring criteria.

Score	Criteria
0–20	All semantics of the translation are unclear or completely wrong, only a small number of words and phrases are correct, and the readability is extremely poor, making it difficult to understand.
21–40	Very few semantics of the translation are correct, key information is missing or wrong, the readability is poor, and a lot of unidiomatic and disfluent expressions and grammatical errors exist.
41–60	Some key semantics of the translation are correct, but a large number of non-key semantics are wrong and the fluency and idiomaticity are poor.
61–80	Most key semantics of the translation are correct, but some non-key information is wrong and major grammatical errors and unidiomatic expressions exist.
81–100	All semantics of the translation are correct, only a little non-key information is wrong, minor grammatical errors exist, and the fluency and idiomaticity are good.

use are developed mainly based on translation accuracy and fluency dimensions specified by China Conference on Machine Translation (CCMT).

2 Evaluation Settings

2.1 Models

We mainly compare the translation results of ChatGPT with Google Translate¹ and Huawei Translate (which is a proprietary engine based on transformer). The results in this paper come from the *gpt-3.5-turbo* models, which power ChatGPT. And the translation prompt is “*Please translate the following sentence into Chinese: <sentence>*”.

2.2 Data

We construct three English-Chinese (en-zh) datasets from one general domain and two specific domains (biomedical and ICT), and choose 200 parallel sentences randomly sampled from the WMT22 General MT Task [4], WMT22 Biomedical MT Task [5], and ICT translation task, respectively.

2.3 Metrics

In this paper, we utilize DA metrics [1] to evaluate ChatGPT and the two translation systems on the three datasets. DA is a commonly used human evaluation method for translation, which involves collecting human assessments of translation quality for single MT systems. It requires highly trained and experienced assessors in the domains concerned. Assessors are provided with a

¹ <https://translate.google.com>

Table 2. DA, BLEU and COMET scores for the three systems (Huawei, Google, and ChatGPT) in the general, biomedical and ICT domains.

System	DA	BLEU	COMET
	<i>General</i>		
Huawei	73.03	48.49	85.3
Google	77.86	49.60	86.7
ChatGPT	77.61	40.64	86.9
<i>Biomedical</i>			
Huawei	77.90	45.71	86.0
Google	80.24	49.25	86.7
ChatGPT	75.39	39.56	85.9
<i>ICT</i>			
Huawei	93.03	32.83	82.4
Google	92.49	33.49	81.9
ChatGPT	84.21	28.12	81.0

candidate translation and a corresponding translation hint (e.g. the source text, a reference translation, or multimodal content) and are asked to assign a quality score from 0 to 100. Our scoring criteria for DA are shown in Table 1. Following the DA criteria, 7 assessors with more than 10 years of translation experience score the 600 sentences separately based on source texts and references. For each sentence, we have at least 3 assessors score it and take the average score as its DA score. In addition, two commonly used automatic metrics BLEU [6] and COMET (wmt22-COMET-da) [8] are used for reference.

3 Evaluation Results

3.1 DA Scores

As shown in Table 2, the DA score for ChatGPT is 77.61 in the general domain, 75.39 in the biomedical domain, and 84.21 in the ICT domain. In the general domain, ChatGPT’s DA score is just 0.25 points ($77.86 - 77.61$) lower than that of the best performer, Google Translate. However, in the ICT domain, its DA score is much lower than those of the other two (more than 5 points lower). The DA results are basically consistent with the BLEU and COMET results. The reason for the inferior performance of ChatGPT in translating domain-specific terminologies is that ChatGPT is trained using publicly available data and is seldom trained in domain-specific terminologies.

As a whole, ChatGPT achieves competitive performance compared with Google Translate and Huawei Translate, although it is not the best among the three systems in any domains covered in this paper. The result aligns with recent research in [7] and [2].

Table 3. A case of idiom translation, where only ChatGPT correctly translates the real meaning of "pull his shit on".

SRC	This is letting Uri Geller try and pull his shit on James Randi.
REF	这是让尤里·盖勒在詹姆斯·兰迪身上 耍花招 。
ChatGPT	这是让Uri Geller在詹姆斯·兰迪身上 耍花招 *。
Google	这是让Uri Geller 试图对James Randi 拉屎 *。
Huawei	这让Uri Geller尝试 把他的屎拉到 *James Randi身上。

* 耍花招 indicates playing tricks; 拉屎 literally means pooping; 把他的屎拉到 indicates putting one's poop onto.

Table 4. A case of ICT terminology translation, where only Huawei Translate correctly translates the terminology "BIER header".

SRC	Multicast packet forwarding in a BIER domain is based on the BitString field in the BIER header .
REF	BIER域中的组播数据包基于 BIER报文头 中的BitString字段转发。
ChatGPT	BIER 域中的组播数据包转发基于 BIER标头 *中的BitString 字段。
Google	BIER 域中的组播数据包转发基于 BIER标头 *中的BitString 字段。
Huawei	BIER域中的组播报文转发基于 BIER报文头 *中的BitString字段。

* BIER报文头 means a BIER header, and BIER 标头 means a BIER label header.

3.2 Case Analysis

As shown in Table 2, in the general domain, the DA score for ChatGPT is very close to that for Google Translate and much higher than that for Huawei Translate. And its COMET score is the highest. This means that ChatGPT performs very well in translation tasks for the general domain. However, ChatGPT gets the lowest BLEU score in this domain. One of the reasons is that BLEU does not reflect the semantic accuracy, and a correctly translated sentence may still receive a low score.

Merits of ChatGPT: A distinctive feature of ChatGPT is that it can produce fluent and natural results on colloquial and informal text. Translationese describes awkwardness in text generated during translation that distinguishes translated text from original ones. ChatGPT suffers the least from translationese among the three evaluated models because it is built on the language training model. For instance, overly literal translation of idioms is one type of translationese. ChatGPT manages to translate beyond literal words and present the real meaning. An example is shown in Table 3, where only ChatGPT correctly translates the idiom "pull his shit on".

Weakness of ChatGPT: While we credit ChatGPT with fluent outputs on colloquial text, its results appear to be informal when it comes to the ICT and biomedical domains, where the text style is distinctively different from that of the general domain. In addition, ChatGPT's performance is inferior to the other two systems in the two domains that involve a lot of domain-specific terminologies. In particular, ChatGPT is sometimes unable to identify the implications of words and as a result produces sentences with poor logic. In comparison, Huawei Translate

is adapted to the ICT domain so it has the greatest terminology accuracy in this domain. As shown in Table 4, "BIER header" should be translated into "BIER报文头" instead of "BIER标头". The result implies that domain adaptation might be the key enabler for standard MT models to compete with foundation models.

4 Conclusion

Regarding English→Chinese translation, ChatGPT as a whole achieves comparable performance with standard MT models, especially in the general domain, where the source text is colloquial and does not involve many terminologies. In this domain, ChatGPT can produce fluent and natural results and suffers the least from translationese among the three evaluated models. However, in the biomedical and ICT domains, where a large number of domain-specific terminologies are involved, ChatGPT is incompetent to translate the terminologies in many cases. It is worth researching on whether ChatGPT's performance can be further improved when external knowledge is input into it. Currently, directly using ChatGPT for domain-specific translation is not recommended.

References

1. Graham, Y., Baldwin, T., Moffat, A., Zobel, J.: Continuous measurement scales in human evaluation of machine translation. In: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. Association for Computational Linguistics, Sofia, Bulgaria (Aug 2013)
2. Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y.J., Afify, M., Awadalla, H.H.: How good are gpt models at machine translation? a comprehensive evaluation. arXiv preprint arXiv:2302.09210 (2023)
3. Jiao, W., Wang, W., tse Huang, J., Wang, X., Tu, Z.: Is chatgpt a good translator? a preliminary study. In: ArXiv (2023)
4. Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., Graham, Y., Grundkiewicz, R., Haddow, B., Knowles, R., Koehn, P., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Novák, M., Popel, M., Popović, M.: Findings of the 2022 conference on machine translation (WMT22). In: Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 1–45. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid) (Dec 2022), <https://aclanthology.org/2022.wmt-1.1>
5. Neves, M., Jimeno Yepes, A., Siu, A., Roller, R., Thomas, P., Vicente Navarro, M., Yeganova, L., Wiemann, D., Di Nunzio, G.M., Vezzani, F., Gerardin, C., Bawden, R., Estrada, D.J., Lima-Lopez, S., Farre-Maduel, E., Krallinger, M., Grozea, C., Neveol, A.: Findings of the wmt 2022 biomedical translation shared task: Monolingual clinical case reports. In: Proceedings of the Seventh Conference on Machine Translation. pp. 694–723. Association for Computational Linguistics, Abu Dhabi (December 2022), <https://aclanthology.org/2022.wmt-1.69>
6. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002)

7. Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., Tao, D.: Towards making the most of chatgpt for machine translation. arxiv preprint (2023)
8. Rei, R., Stewart, C., Farinha, A.C., Lavie, A.: COMET: A neural framework for MT evaluation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 2685–2702. Association for Computational Linguistics, Online (Nov 2020), <https://aclanthology.org/2020.emnlp-main.213>

Comparing MTQE Scores with Fuzzy Match Percentages from CAT Tools

Elena Murgolo¹

¹ Orbital14, Milano, Italy
emurgolo@orbital14.ai

Abstract. Due to the increasing integration of machine translation (MT) workflows in the day-to-day operations of many LSPs around the world, the raw output quality is becoming one of the key elements to be considered during translation and localization tasks, as it has a direct influence on turn-around times and overall costs. One of the key factors in the use of machine translation quality estimation (MTQE) tools will be the possibility to create readable reports for all stakeholders within the localization workflow. In the following, a preliminary work on the analysis of MTQE scores and on the creation of such a report by comparing them with fuzzy match percentages from CAT tools will be illustrated.

Keywords: MTQE, Quality Estimation, Fuzzy Matches.

1 Introduction

In today's machine translation (MT) workflows, the quality of the MT raw output is usually evaluated after a post-editing (PE) step to be carried out by a linguist or bilingual professional. The MT is compared to the version of the text that has been reviewed and corrected by a human and different quality metrics and comparison methods can be used to evaluate how good the original translation was. Typical such metrics include, for instance, BLEU scores [5], TER [8], chrF++ [7], etc. All these evaluation methods rely on a human reference and are therefore necessarily carried out at the end of the translation process, after PE.

MT evaluation is essential to monitor the MT engines' performance and to manage any retrain and/or finetuning steps on the models.

Contrary to MT evaluation, machine translation quality estimation (MTQE) models give an automatic estimate of the quality of the MT raw output without relying on human reference texts and may, therefore, be deployed at the beginning of the process for a better overview of costs, turn-around times, PE effort, and/or to determine which MT engine is better suited for the content at hand.

Depending on the single use cases, different kind of metrics can be used as the MTQE models' output, and they can be displayed in different ways to the end user based on each case's needs.

One popular such method of displaying the result consists in showing the MTQE analysis results to the post editors with a “traffic light” coding [1, 6, 10] representing a binary indication of the estimated quality of the MT raw output. The linguist in this case is given an indication such as:

- Good/green: the MT output can be used for post-editing according to the automatic estimation as it requires little to no changes;
- Bad/red: the MT suggestion should be discarded as a human translation from scratch would require less time/effort.

Béchara et al. [1] based their traffic light system on fuzzy match scores as “translators are more used to working with TM leveraging and fuzzy matches” and used a 75% quality threshold to predict the good/bad evaluation. In their study, however, fuzzy match percentages are not shown to the post-editors who are then provided with the quality estimation by means of color codes.

It is worth mentioning that previous testing in user settings has shown contradicting results on the usefulness of displaying MTQE scores to post-editors [1, 6, 9, 10], the interest for MTQE is still rising there have been suggestions for the fact that MT confidence score may be useful during translation and localization workflows and Moorkens et al. report that most post-editors who took part in their survey would also like “to be presented with confidence scores for each target text segment from the MT engine.” [3].

In the following sections, the possibility will be discussed to find a correspondence between fuzzy match percentages and TER scores resulting from MTQE, with the aim of finding an alternative way to present MTQE results to end-users. The analysis is still ongoing, and this paper will present only preliminary considerations on the matter.

2 MTQE Tool

The tool presented in the following sections is part of a machine translation quality assessment (MTQA) product developed to be used in production [4].

The MTQA application was designed as a standalone tool including MT quality evaluation metrics and MTQE models to be used by non-programmers in a localization workflow.

2.1 Standalone MTQA Tool

The MTQA tool was developed as a standalone application so it could address the need for flexibility required for typical localization and translation workflows within language service providers (LSPs), where the MTQA tool will need to be integrated with other existing applications using different file standards.

It is nowadays not uncommon for LSPs to deploy different computer-assisted (CAT) tools during their workflows and even different MT engines, either trained on specific domains or generic engines. Such MT models may as well be supplied by different providers depending on language combinations, price, trainability, and other factors.

A standalone tool would allow users to obtain an MTQE analysis of any job regardless of the different systems used to manage the files.

2.2 MTQE Model

The results taken into consideration for this paper come from an MTQE model based on Google’s BERT [2] and trained with proprietary industry data consisting of a source text in English, the MT raw output in Italian and a post-edited version, in the technical domain. The labelling for the training consisted of the TER scores between the MT and the PE at a sentence-by-sentence level.

The model provides a hypothesis on how many changes are needed to change the MT output into the closest possible correct version.

The results of the MTQE are presented to the end user in a .csv file by providing an estimation of the edit distance for each sentence in the analysed MT raw output text.

index	SRC	MT	predictions
0	SLIDE VALVES	VALVOLE PER SLITTA	41.11102670431140
1	BACKGROUND OF THE INVENTION	STATO DELL'ARTE DELL'INVENZIONE	21.079684046891100
2	1. Field of the Invention	1. Campo dell'invenzione	18.704068288207100
3	The present invention relates to slide valves for fluid cata	La presente invenzione riguarda valvole per slitta per unità di cracking c	46.53065353631970
4	2. Background and Related Art	2. Stato dell'arte e arte correlata	36.88893020153050
5	The actuator is a key component for a fluid catalytic crack	L'attuatore è un componente chiave per un funzionamento e una presta	41.90143793821340
6	Existing actuators continue to have certain deficiencies. In	Attuatori esistenti continuano ad avere alcune mancanze. In molti casi, l	29.062977433204600
7	Some existing actuators include a manual actuation elem	Alcuni attuatori esistenti includono un elemento di azionamento manual	27.54213809967040
8	JIP2004 257487 discloses a half nut device or split nut de	JP 2004 257487 divulga un dispositivo per semidado o un dispositivo di	30.43023347854610

Fig. 1 Example of a .csv file with sentence-by-sentence estimation scores.

2.3 Interpretation of Results

For the tool to be usable in a production setting, the aforementioned list of sentence-by-sentence numbers needs to be easily readable by people working in difference capacities within an LSP or localization department. On the one hand, project managers will need a quick overview of the MT quality to manage costs and time. On the other hand, linguists should be able to read the results to decide whether to discard the MT output as the quality will be too low or to use it as a starting point for PE, and, in the latter case, to gauge the necessary effort for each segment.

The approach chosen to create a readable report for project managers and linguists was to compare the MTQE results for each sentence to a hypothetical fuzzy band match between the MT raw output and the post-edited version.

Fuzzy Matches Scores. To create the fuzzy match scores for each sentence, mock projects are set up on common commercial computer-assisted translation (CAT) tools. Translation memories (TMs) are created ad-hoc, in which the MT raw output is used as source and the text in the source language as target. When the mock project is created, the post-edited version is used as source text and a TM match batch task is launched using the lowest possible pretranslation threshold.

This will create a pretranslated bilingual file containing fuzzy match percentages calculated between the two Italian versions (MT and PE).

It is important to take into consideration that fuzzy match percentages of the same sentence do not necessarily correspond from one CAT to another. In fact, in a sample text containing 9254 words and 528 segments, 70% of the segments did not receive the same scoring, and 55% of the total number of segments would not be categorized in the same fuzzy match band in a common CAT fuzzy analysis, either.

Table 1. Analysis of an example text by two common commercial CAT tools. The fuzzy matches scores are calculated by comparing raw MT with its post-edited version.

	CAT 1		CAT 2	
	Word count	Words %	Word count	Words %
100%	1037	11%	1111	12%
95%-99%	377	4%	2061	22%
85%-94%	2410	26%	4057	44%
75%-84%	2825	31%	1336	14%
50%-74%	2270	25%	657	7%
New	335	4%	32	0%
Total	9254		9254	

Skewed Results. During the analysis it was clear that the distribution of both the fuzzy match percentages and of the MTQE scores was strongly skewed, showing a majority of good to very good quality segments, either with very low TER scores from MTQE or with very high fuzzy indication.

This may show that the training data used to create the model were themselves biased, the MT raw output used to create the TER labels being of too high quality and not containing enough low-quality segments for the engine to learn correctly.

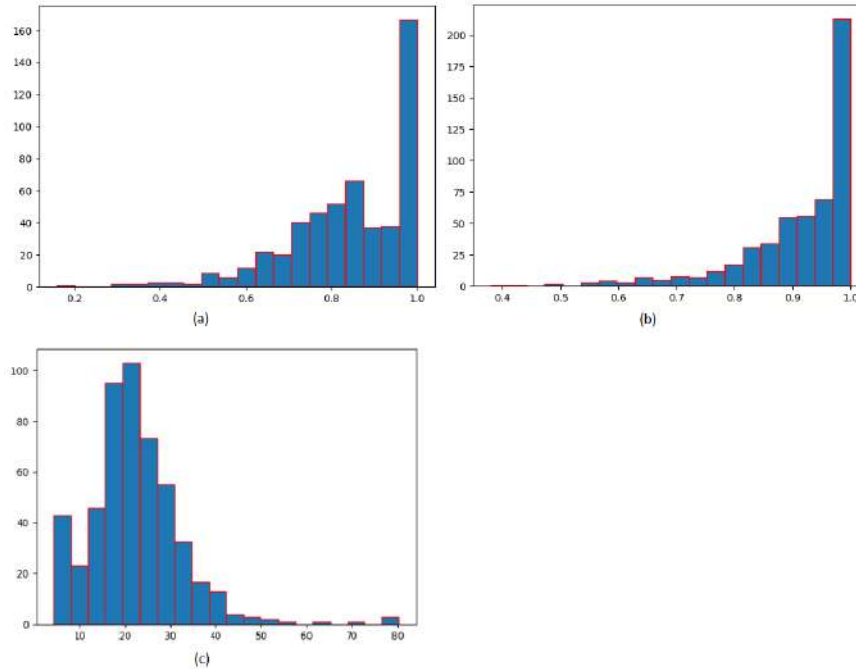


Fig. 2 Distribution graphs for (a) fuzzy matches from CAT1, (b) fuzzy matches from CAT2, and (c) MTQE scores (expressed in TER, where the lower the score, the less changes should be applied and therefore the higher the quality).

MTQE-Fuzzy Matching. The matching process to create an MTQE analysis to be used in production is still on-going.

The skewed distribution of the CAT fuzzy match percentages has a direct influence on the availability of segments in the lower bands to be compared with the MTQE scores, thus slowing down the mapping of a possible QE analysis.

3 Conclusions and Future Work

The one presented above is only one of the possible approaches both for the creation of an MTQE tool and for the development of user-readable quality reports.

Future work for the continuation of the analysis described above includes the re-training of the models with more symmetrical data and the use of comparison texts containing a higher number of fuzzies within the lower bands (50%-74% and 0%-49%).

Though clarity in the interpretation of results from MTQE models will be a key factor in the uptake of such tools, there can be numerous approaches to obtaining it, only one of which has been presented herein, which may be useful for some use-cases and not applicable for others.

References

1. Béchara, H., Orăsan, C., Parra Escartín, C., Zampieri, M., and Lowe, W.. *The Role of Machine Translation Quality Estimation in the Post-Editing Workflow*. Informatics 2021, 8, 61.
2. Devlin J., Chang M., Lee K., and Toutanova K.. 2018. *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv:1810.04805
3. Moorkens, J., and O'Brien, S.. *Assessing User Interface Needs of Post-Editors of Machine Translation*. In *Human Issues in Translation Technology: The IATIS Yearbook*; Kenny, D., Ed.; Routledge: Oxford, UK, 2017; pp. 109–130
4. Murgolo E., Pourmostafa Roshan Sharami J., and Shterionov D.. 2022. *A Quality Estimation and Quality Evaluation Tool for the Translation Industry*. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 307–308, Ghent, Belgium. European Association for Machine Translation.
5. Papineni K., Roukos S., Ward T., and Zhu W.. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.
6. Parra Escartín, C., Béchara, H., and Orăsan, C.. *Questing for Quality Estimation A User Study*. *Prague Bull. Math. Linguist.* 2017, 108, 343–354.
7. Popović M.. 2017. “*chrF++: words helping character n-grams*”. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
8. Snover M., Dorr B.J., Schwartz R., Makhoul J., Micciulla L., and Weischedel R.. 2005 “*A Study of Translation Error Rate with Targeted Human Annotation*”, LAMP-TR-126, CS-TR-4755, UMIACS-TR-2005-58, University of Maryland, College Park, MD.
9. Teixeira, C., and O'Brien, S.. *The Impact of MT Quality Estimation on Post-Editing Effort*. In *Proceedings of the MT Summit XVI, Volume 2: Users and Translators Track*, Nagoya, Japan, 18–22 September 2017; pp. 211–233.
10. Turchi M., Negri M., and Federico M.. *MT Quality Estimation for Computer assisted Translation: Does it Really Help?* In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 530–535, Beijing, China, July 26–31 2015. Association for Computational Linguistics.

Are post-edited features really universal?

Lise Volkart¹ and Pierrette Bouillon¹

University of Geneva - lise.volkart|pierrette.bouillon@unige.ch

Abstract. This study aims at assessing whether or not differences between human and post-edited machine translations, so-called post-edited features, are common to all post-edited texts. We investigate the presence of four candidate post-edited features in corpora that include two language pairs and three different machine translation systems to find out whether or not different post-edited texts share the same characteristics, regardless of other factors. Our results show that the language pair, as well as the MT system, can have a great influence on these four features, thereby going against the assumption that they are common to all post-edited texts.

Keywords: Post-edited · Post-editing · Human translation · Universal features · Machine translation.

1 Introduction

With the recent advances of neural machine translation (NMT), numerous language service providers and professional translators are adopting machine translation post-editing as a new way of working. As a result, we are likely to be increasingly exposed to machine-generated texts that are post-edited by human translators. This major shift in the translation sector comes with its share of questions. Some of them, such as the ones regarding the quality of machine translation (MT) and post-editing (PE), have been widely studied, while others, like questions of style, have garnered less attention. By studying the specific characteristics of post-edited machine translations (PEMT), we hope to shed light on these relatively unexplored areas. We focus on so-called post-edited features and raise the question of their universal character.

We start by presenting the motivation and goal of our paper. We then summarise our methodology and introduce our research question, corpora and metrics, before presenting our results. We end with some concluding remarks and perspectives for future work.

2 Motivation and Goal

In recent years, numerous studies have attempted to analyse the characteristics of PEMT in comparison with human translations (HT), employing various methodologies and focusing on different phenomena, such as cognate translations

or terminological variations [14], frequency of certain words or phrases [10], and variation of translation solutions [9] (a more detailed description of these studies and their primary results can be found in [13]). While these studies have yielded promising results in terms of developing our understanding of post-edited language, it remains difficult to generalise on the characteristics of post-edited texts, due the variety of study designs and features under investigation. In an attempt to establish a stronger theoretical and methodological framework for the study of PEMT in contrast with HT, some researchers [3, 4, 7, 11] have decided to examine the question through the prism of corpus-based translation studies (CBTS) and the well-known concepts of translationese and translation universals (also sometimes called “universal features of translation”), as introduced by [1]. In 2017, [7] put forward the notion of post-editese and defined it as “the expected unique characteristics of a post-edited text that set it apart from a [human] translated text”. Following this work, several researchers took part in the quest to identify the universal features of post-edited texts (see [3, 4, 11, 13]).

These studies examine the metrics related to lexical richness or diversity, lexical density and sentence length, among other features. [7] was unable to prove the existence of post-editese, whether based on human judgement or on automatic classification. On the other hand, [3, 4, 11, 13] have observed certain significant differences between HT and PEMT. In [11], the investigated PEMT corpora exhibit lower lexical richness and lexical density than their HT counterparts, as well as a sentence length that is closer to the source (with the exception of one corpus). The hypotheses put forward by [4, 3] are in line with the work of [11], but were not confirmed by all datasets. Similar hypotheses were also advanced by [13], with results confirming the conclusion drawn by [11] with regard to lexical richness and density, but not sentence length.

We identified 3 main axes around which we can work to gain a deeper understanding of these phenomena. First, there is a quantitative axis. Since this area of research is still in its infancy, studies are limited in numbers and scale, making it difficult to generalise. In particular, there is a lack of studies on large datasets. We therefore need large and reliable corpora of HT and PEMT in order to study post-editese. Second, there is a methodological axis. A clear methodological framework for the study of post-editese still has yet to be defined. Corpus type and size vary, as well as the metrics that are being used, which can be very sensitive to independent factors and calculation methods [2]. It is therefore almost impossible to compare results between different studies. The question of the type of corpus used (parallel or comparable) is also a determining factor in post-editese research (for a discussion of this topic, see [12, 13]). Implementing a clear methodological framework would clearly help advance studies in the field. Finally, the third axis is related to the angle of study. The focus on finding systematic differences between texts based solely on the mode of translation (HT vs PEMT) tends to take attention away from other possible factors. This is a problem that has already been identified in CBTS in relation to the study of translation universals [8]. As language is a multi-factorial product, it can be influenced by a number of factors, and each of these factors, or combinations of them,

can account for some of the differences observed. The influence of other potential factors (such as language pair, MT system, domain, and experience/status of the post-editor) is often neglected in favour of the translation mode. Furthermore, the ultimate goal of identifying “universal” post-edited features diverts the attention of researcher from post-edited features that are specific to certain language pairs, domains and MT systems. We are therefore convinced that the field would benefit from adopting a broader perspective.

With regard to the third axis, the aim of our work is to take a step back from previous studies and interrogate the universal character of previously identified post-edited features by comparing them across different corpora. We also attempt to address the first and second axes by working with several relatively large corpora and applying the same methodology to all of them.

3 Methodology

3.1 Research Question

In this study, we compare HT and PEMT on different corpora with different language pairs and different MT systems, in order to answer the following research question:

Are we able to identify consistent features in post-edited across corpora with different source languages, different domains and different MT systems?

In other words, post-edited features are examined while taking into account not only the translation mode, but also the corpus, source language and MT system as independent variables.

3.2 Corpus

Corpus design: Our experiment is based on two distinct corpora made up of authentic translation jobs gathered from two different companies¹. The design of both corpora is identical and the same methodology was used to compile them. CorpusDEfr contains professional translations from German into French gathered from an insurance company based in Switzerland. CorpusENfr also contains professional translations, but from English into French, which originate from a sports organisation, also based in Switzerland.

Both corpora consist of two sub-corpora of the same size, one made up of PEMT segments and the other made up of HT segments. For each sub-corpora, we have access to source and target segments². For CorpusENfr, PEMT segments were all produced using the same MT system (a customized NMT engine), while

¹ For a discussion of the advantages and limitations of working with authentic data, see [13].

² For PEMT, we only have access to the final post-edited segments, as raw MT is not saved during the translation process.

for CorpusDEfr, PEMT segments were produced using three different generic NMT systems. For each language pair, HT and PEMT sub-corpora are assumed to be comparable. Indeed, they contain the same type of texts, translated in the same context and by the same team of translators.

Corpus creation methodology: The first step in corpus creation consists of gathering human and post-edited translation material. Data was gathered in the form of translation memories and the translations were all produced by professional translators in their usual working environment.

Once the material had been gathered, the following pre-processing steps were applied:

- Extraction of the translated content (source and target, segment by segment)
- Identification and subdivision of HT and PEMT segments
- NER (Named entity recognition) tagging and anonymisation (only for the CorpusDEfr)³
- Cleaning steps (removing urls, non alphabetical segments, potential personal information, etc.)
- Identification and deletion of duplicate segments
- Sampling by source segment length (removing segments shorter than 6 tokens and longer than 40 tokens)
- Random sampling to obtain the same size for both translation modes (HT and PEMT).

Through this pre-processing, we seek to attain clean and comparable sub-corpora (HT and PEMT) for each corpus/language direction. Figures 1 and 2 present the structure of each corpus as well as the number of tokens for each corpus and sub-corpus.

3.3 Metrics

Below, we present, the four different metrics selected for this experiment. The first three (or variants of them) are commonly used to identify post-edited. To the best of our knowledge, the fourth metric has never been used to investigate post-edited.

Type token ratio variation between source and target (TTRvar): Type-token ratio (TTR) is a metric that is commonly used in the study of post-edited, it reflects the lexical richness of a text or corpus. It is computed by dividing the number of unique words by the total number of words [2]. Given that HT and PEMT target corpora are derived from different source corpora, target TTR scores cannot be directly compared and we have to take into account potential differences in TTR scores between the sources [13]. We therefore decided to compute and compare the percentage of variation between source and target in

³ The anonymisation was required by the corpus provider.

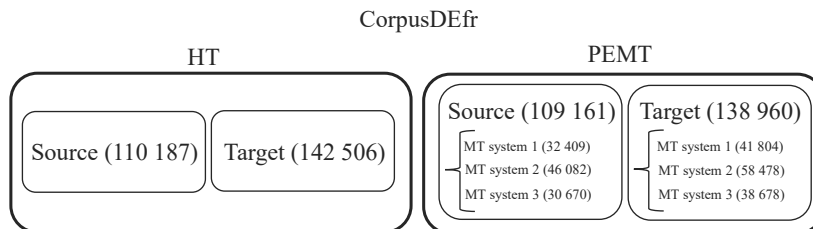


Fig. 1. CorpusDEfr structure. Number of tokens for each corpus/sub-corpus is indicated between parentheses.

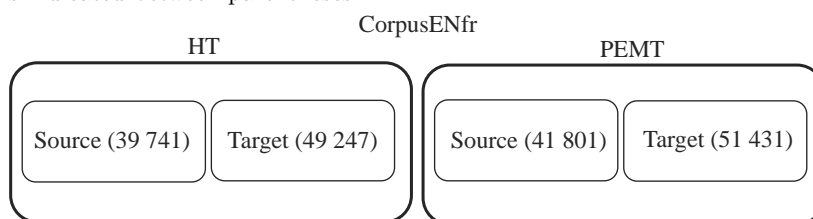


Fig. 2. CorpusENfr structure. Number of tokens for each corpus/sub-corpus is indicated between parentheses.

both translation modes and refer to this score as TTRvar. The TTRvar score is computed as follows:

$$TTRvar = \frac{TTR_{target} - TTR_{source}}{TTR_{source}} \times 100$$

A negative TTRvar means that, for the given language pair, the target exhibits a lower TTR than its source counterpart.

Lexical density variation between source and target (LDvar): Lexical density corresponds to the number of lexical words (verbs, adverbs, nouns, proper nouns and adjectives) divided by the total number of words [11]. Post-edite studies often include this score as a measure of the amount of information contained in a text [11, 4]. Again, the influence of the source corpora has to be taken into account when comparing lexical density. Hence, we compute the percentage of variation of LD between source and target (LDvar), based on the same formulae as for TTRvar. Here again, a negative score indicates that the target has a lower LD than its source counterpart.

Expanding ratio (ER): Measurements related to sentence length are also often part of post-edite studies, either in the form of mean sentence length [3], length ratio [11] or expanding ratio [10, 13]. We opted for the latter, given that

it is a percentage that reflects the sentence length variation between source and target, as measured in words [6, 5], thus fulfilling the need to take into account source variability. It is computed using the following equation:

$$ER = \frac{Length_{target} - Length_{source}}{Length_{source}} \times 100$$

A positive ratio means that the target segment is longer than the source.

Parsing height variation (Parsvar): The parsing height corresponds to the maximum height of the dependency parsing tree at sentence level. It is computed by counting the maximum number of heads of each sentence’s dependency parsing tree. Parsvar score corresponds to the percentage of variation in parsing height between source and target. It is computed in the same way as TTRvar and LDvar. We consider the Parsvar score to be an indicator of sentence structure similarity between source and target. We assume that when Parsvar is close to zero (that is, the parsing tree of source and target are of the same height), source and target are more likely to have similar syntactic structures.

Metrics are computed for both corpora and for each sub-corpora (HT and PEMT). Segments are shuffled (to avoid the influence of different topics) and grouped into chunks of 75 sentences. TTRvar and LDvar are computed at chunk level and ER and Parsvar are computed at sentence level and averaged on chunk level. All metrics are computed using ad hoc python scripts. Parsing and tagging are performed using SpaCy’s English, French and German small models⁴.

4 Results

4.1 Results across corpora

In this subsection, we compare the results obtained for CorpusDEfr and CorpusENfr. Our goal is to establish whether or not consistent signs of post-editease can be found across corpora with different source languages.

TTRvar: Figure 3 presents TTRvar scores for HT and PEMT for CorpusDEfr and CorpusENfr. Scores are negative for all corpora and sub-corpora, indicating that the translation step results in a drop of TTR scores for both language pairs and translation modes. We also remark that with both corpora, the drop in TTR scores is less important for PEMT than for HT (however, this difference is only significant for CorpusDEfr ($p < 0.001^5$)).

⁴ <https://spacy.io/models>

⁵ For all metrics, statistical significance was tested using the Mann-Whitney non-parametric test, as data is not normally distributed. Extreme outliers were removed for plotting and significance testing

TTRvar differences between HT and PEMT are consistent across our corpora and reveal a tendency toward higher lexical richness in PEMT, as shown by the lower drop in TTR score. It is worth noting that this result contradicts what has been observed in previous studies, where the loss of lexical richness is generally shown to be more important in PEMT [11, 12].

LDvar: Similarly to TTRvar, the LDvar scores presented in Figure 4 indicate a tendency toward a loss of lexical density during translation for both corpora and translation modes. For CorpusDEfr this loss is more important in PEMT than HT (significant at $p < 0.05$), which is in line with the findings of previous studies and supports the hypothesis that PEMT produces simpler translations [11]. However, this tendency is not confirmed for CorpusENfr, where the loss of LD is more important in HT.

In terms of LDvar, the differences observed between HT and PEMT are not consistent across corpora.

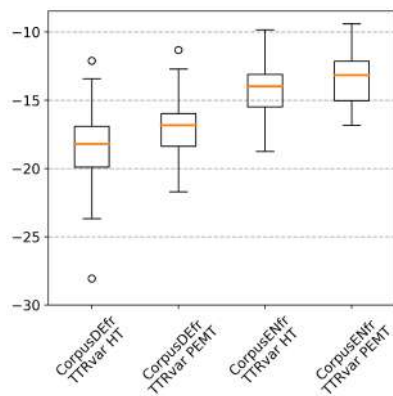


Fig. 3. TTRvar scores for HT and PEMT for CorpusDEfr and CorpusENfr

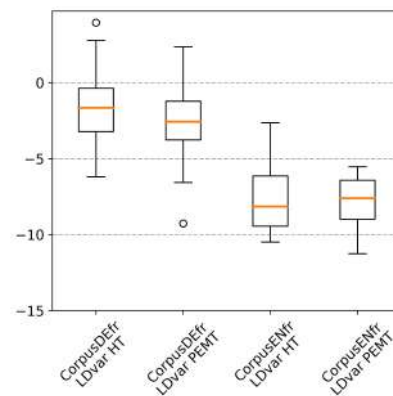


Fig. 4. LDvar scores for HT and PEMT for CorpusDEfr and CorpusENfr

ER: Figure 5 presents the ER scores for both corpora. Scores are all positive, which indicates that target segments are longer than their source equivalent in both corpora. In CorpusDEfr, PEMT exhibits a lower ER than HT, meaning that for a given source segment length, the target tends to be shorter in PEMT (significant at $p < 0.001$). This tendency is not confirmed in CorpusENfr, where no significant difference between HT and PEMT can be detected in terms of ER.

Here again, the differences that are observed between HT and PEMT do not follow the same tendency in both corpora.

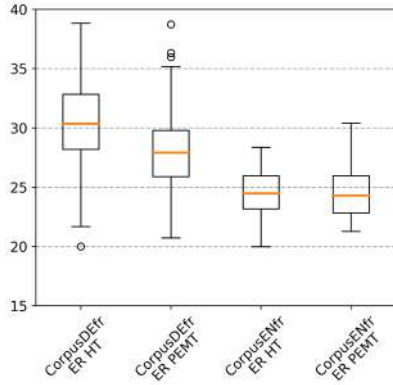


Fig. 5. ER scores for HT and PEMT for CorpusDEfr and CorpusENfr

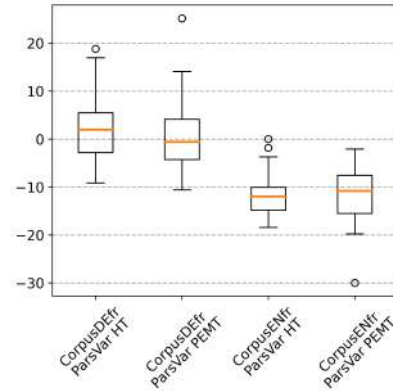


Fig. 6. ParsVar scores for HT and PEMT for CorpusDEfr and CorpusENfr

Parsvar: Figure 6 presents the Parsvar scores for HT and PEMT for CorpusDEfr and CorpusENfr. With medians close to zero for both translation modes and a large spread in the scores, our German-into-French corpus does not exhibit any significant difference between HT and PEMT, in terms of sentence structure similarity between source and target. For corpusENfr the distribution of the Parsvar scores is also not significantly different between HT and PEMT, but scores are less spread and the entire population is situated below zero. If, as hypothesised, a Parsvar close to zero is an indication of sentence structure similarity between source and target, this means that sentence structure varies more in CorpusENfr than CorpusDEfr.

These results indicate the strong influence of the corpora and the source language on the parsing height of sentences. On the contrary, translation mode does not seem to affect this metric.

4.2 Results across MT systems

In the second part of our analysis, we divide CorpusDEfr PEMT into three parts, according to the MT system used (System 1, System 2, System 3). We compute the same scores for each part and compare results. Our goal is to verify whether consistent signs of post-editeuse can be identified across different MT systems.

TTRvar: In Figure 7, we can see that PEMT presents a greater loss of lexical richness when using System 1 than System 2 ($p < 0.005$) and System 3 ($p = 0.005$). The differences suggest that the MT system has an influence on the lexical richness of the final output in PEMT.

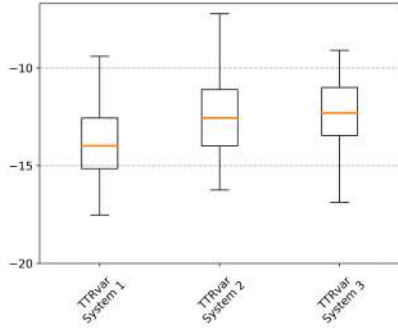


Fig. 7. CorpusDEfr TTRvar scores for each MT system.

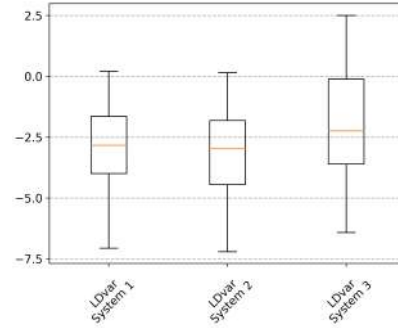


Fig. 8. CorpusDEfr LDvar scores for each MT system.

LDvar: LDvar scores across MT systems are presented in Figure 8. The loss of lexical density is more limited for System 3 than for Systems 1 and 2. The difference between Systems 2 and 3 is statistically significant at $p=0.01$. The MT system used for PEMT therefore also seems to have an influence on the final output in terms of lexical density.

ER: As for the ER scores presented in Figure 9, System 1 exhibits a slightly higher score than Systems 2 and 3, but no statistically significant difference was found between systems in terms of ER. The MT system does not seem to have an influence on the length ratio of the final output.

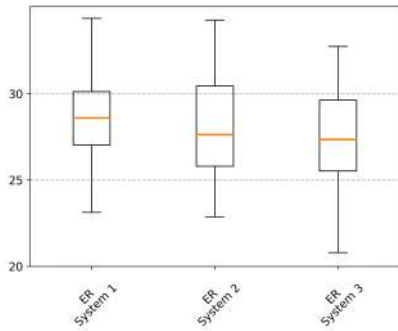


Fig. 9. CorpusDEfr ER scores for each MT system.

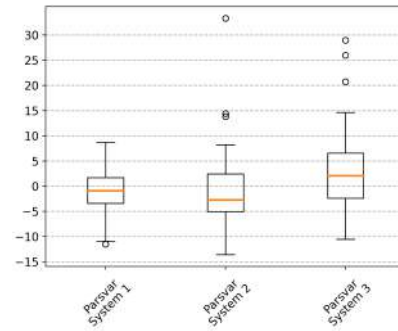


Fig. 10. CorpusDEfr Parsvar scores for each MT system.

Parsvar: Parsvar scores for each MT system are presented in Figure 10. We note that the Parsvar score is closer to zero for System 1 than for the two others. Based on our hypothesis on Parsvar score, this means that the use of Systems 2 and 3 leads to more sentence structure variation between source and target. On the contrary, System 1 leads to less variation. The distribution of scores for System 3 is significantly different from System 1 at $p=0.05$ and from System 2 at $p<0.05$, which suggests that the MT system has an influence on the sentence structure of the final output.

5 Conclusion

The aim of this experiment was to study the consistency of four candidate post-edite features (lexical richness, lexical density, length ratio and sentence structure similarity between source and target) across different corpora, source languages and MT systems, in order to assess their credibility as universal features of post-edite.

With only one exception, we were unable to identify consistent features of post-edite across corpora with different source languages and different MT systems.

The comparison across corpora shows that these post-edite features are not constant and are therefore influenced by the corpus and/or source language. The only shared characteristic between corpora is a tendency toward higher lexical richness in PEMT, but this contradicts the results of previous studies.

The comparison across MT systems reveals that the system also has an influence on post-edite features. Differences were observed between systems, in terms of lexical richness, lexical density and sentence structure similarity. Only the feature related to sentence length was consistent across all systems.

Our results do not support the idea of there being universal features common to all PEMT texts. With respect to the features under investigation, we have shown that the influence of the corpora, language pairs and MT system is often greater than the translation mode.

This paper highlights the importance of viewing post-edite as a sample of possible phenomena, rather than a universal set of features shared by all PEMT texts. A large number of factors should be taken into account when looking for differences between HT and PEMT and not the translation mode alone. In addition to the corpus, language pair and MT system, these factors may also include the translator/post-editor's experience/attitude towards MT, text type, etc.

For future work, we intend to continue in the same direction and focus on describing specific post-edite features, while taking into account as many variables as possible. We also aim to expand on the sample of investigated features, in particular by focusing on non-lexical features that remain under-studied in the field.

Acknowledgements We would like to thank the organisations and language services that agreed to share their data with us.

References

1. Baker, M.: Corpus Linguistics and Translation Studies: Implications and applications. In: Baker, M., Francis, G., Tognini-Bonelli, E. (eds.) *Text and Technology. In Honour of John Sinclair*, pp. 233–250. Routledge (1993)
2. Brezina, V.: *Statistics in corpus linguistics: A practical guide*. Cambridge University Press (2018)
3. Castilho, S., Resende, N.: Post-Editese in Literary Translations. *Information* **13**(2), 66 (2022), publisher: Multidisciplinary Digital Publishing Institute
4. Castilho, S., Resende, N., Mitkov, R.: What Influences the Features of Post-editese? A Preliminary Study. In: *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*. pp. 19–27. Varna, Bulgaria (2019)
5. Cochrane, G.: Le foisonnement dans les textes de spécialité, illusion d’optique ou réalité quantifiable?. (2000)
6. Cochrane, G.: Le foisonnement, phénomène complexe. *TTR: traduction, terminologie, rédaction* **8**(2), 175–193 (1995), publisher: Association canadienne de traductologie
7. Daems, J., De Clercq, O., Macken, L.: Translationese and Post-editese: How comparable is comparable quality? *Linguistica Antverpiensia* pp. 89–103 (2017)
8. De Sutter, G., Lefer, M.A.: On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives* **28**(1), 1–23 (2020), publisher: Taylor & Francis
9. Farrell, M.: Machine Translation Markers in Post-Edited Machine Translation Output. In: *Proceedings of the 40th Conference Translating and the Computer*. pp. 50–59. London (2018)
10. Martikainen, H., Kübler, N.: Ergonomie cognitive de la post-édition de traduction automatique : enjeux pour la qualité des traductions. *ILCEA* (27) (Nov 2016). <https://doi.org/10.4000/ilcea.3863>
11. Toral, A.: Post-editese: an Exacerbated Translationese. In: *Proceedings of MT Summit XVII. vol. 1*, pp. 273 – 281. Dublin, Ireland (2019), arXiv: 1907.00900
12. Volkart, L., Bouillon, P.: Building a corpus for the study of post-editese in professional context: methodology and challenges. In: *Book of Abstracts*. p. 195. Prague, Czech Republic (2022)
13. Volkart, L., Bouillon, P.: Studying Post-Editese in a Professional Context: A Pilot Study. In: *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. pp. 71–79. Ghent, Belgium (2022)
14. Čulo, O., Nitzke, J.: Patterns of Terminological Variation in Post-editing and of Cognate Use in Machine Translation in Contrast to Human Translation. In: *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*. pp. 106–114. Riga, Lettonie (2016)

Behind the scenes: Freelance Translators Use of Machine Translation in the Medical Field

Magali Vidrequin^[0000-0002-7676-7377]

Université Rennes 2, Rennes, France
vidrequin.magali@univ-rennes2.fr

Abstract. The increasing adoption of machine translation (MT) in the translation industry has raised questions about the quality and accuracy of MT output, and the role of human post-editing in ensuring the final quality of translated content, especially as a result of the latest developments in Neural Machine Translation. In this study, we explore the attitudes and practices of medical translation professionals with respect to MT and post-editing. A survey was conducted among a sample of 21 freelance medical translators to investigate the extent and forms of MT usage in medical translation, the perceived benefits and challenges of using MT. Our findings highlight the need for ongoing education and training in best practices for data privacy and protection, and the importance of developing tools and strategies for improving the quality and efficiency of medical translation in the age of machine translation.

Keywords: Machine Translation · Freelance Translators · Medical Translation.

1 Introduction

The year 2018 is often cited as a turning point for the use of machine translation (MT) in the translation industry, with more than 50% of language service providers, including freelance translators and translation agencies, reporting some form of MT usage at that time according to a survey conducted in 2018 by EUATC[1]. However, while the growing adoption of MT has been widely acknowledged, questions remain about the specific forms and extent of MT usage among translation professionals. In particular, it is unclear what percentage of translation projects processed by each translator involve the use of MT. As such, it is important to continue exploring the attitudes and practices of translation professionals with respect to MT and post-editing.

Although MT offers advantages such as speed and affordability, it has also raised concerns about the quality of translation and the job security of professional translators. To investigate this issue, we conducted semi-structured interviews with freelance translators to explore their views and their use of MT in their work. Our aim is to provide insights into the challenges and opportunities presented by MT from the freelancers' point of view, drawing on the perspectives of professionals who represent a significant portion of the translation workforce.

The field of medical translation is one that is particularly well-suited for exploring the potential benefits and challenges of using MT and post-editing. Medical translation requires a high level of accuracy and specialized knowledge, making it an ideal test case for evaluating the quality and effectiveness of MT output. Moreover, the demand for medical translation services is growing rapidly, driven by factors such as globalization, cross-border collaborations, and the increasing availability of medical information online. This trend has created a need for faster, more cost-effective translation solutions, making MT an increasingly attractive option for medical translation professionals. By examining the attitudes and practices of medical translation professionals with respect to MT and post-editing, we can gain valuable insights into the use of this technology in a highly specialized and demanding field, and inform the development of new strategies for improving the quality and efficiency of medical translation.

After gathering preliminary data on the use of machine translation by freelance medical translators through a questionnaire[8], the following research questions guided our investigation:

1. How do freelance medical translators use MT in their work, and what factors influence their decision to use or not use it?
2. What are freelance medical translators' opinions on the quality of MT output for medical texts, and how does this affect their workflow and productivity?

These questions were designed to explore different aspects of freelance medical translators' engagement with MT and their views on the technology's impact on their daily work. By achieving these objectives, we aimed to provide a nuanced understanding of the benefits and challenges of MT for freelancers translators of a specialized field.

2 Related Work

In the field of translation studies, it is imperative to acknowledge the context in which translators operate and progress, particularly in relation to pragmatic translation. This consideration gains even greater significance when addressing the challenges posed by MT. Numerous scholarly investigations endeavor to examine the prospective advantages of MT; however, the complexities of training these systems can render them unattainable for smaller translation enterprises or individual translators[2][10]. These professionals may lack the required technical skills to devise or train specialized translation algorithms, in addition to confronting the demands associated with hardware performance, which must be capable of supporting the essential process of training corpora for these engines, knowing that access to sizeable parallel text corpora is a fundamental requirement for the efficient development of MT systems.

Machine translation engines are now widely available through plugins or APIs, and can easily be integrated into computer-assisted translation (CAT) tools. This integration has eliminated much of the initial complexity and cost associated with developing a machine translation engine from scratch and has

greatly simplified the process of incorporating machine translation technology into translation workflows for freelance translators themselves.

Kanavos[3] endorses this integrated approach to translation tools, combining the strengths of both translation memories (TM) and MT systems. By leveraging the advantages of the two technologies, Kanavos argues that productivity can be significantly increased: the integration of TM and MT allows for more accurate and efficient translations, leading to a streamlined workflow and improved overall output.

According to Zaretskaya[10], there are two primary types of integration between MT and translation memory systems: internal integration and external integration. Internal integration refers to the seamless incorporation of MT within the translation memory environment, allowing translators to access both resources concurrently, thereby streamlining the translation process. In contrast, external integration entails the use of MT and translation memory systems as separate tools, with the translator manually importing and exporting data between the two systems. This distinction highlights the varying degrees of interconnectedness between these technologies and their potential impact on the efficiency and overall user experience for translators in their daily work.

Reinke[6] builds upon Zaretskaya's categorization of machine translation integration by further dividing external integration into two subcategories. The first method involves batch processing, where segments of the source text that do not yield a high-percentage match with the TM database are processed by MT tools. The second method is interactive processing, in which translators work closely with the MT system, constantly receiving real-time suggestions and making adjustments to ensure accurate translations.

In a recent paper by Quintana and Castilho[5], a comprehensive review was conducted comparing various CAT tools and their integration of MT capabilities according to several criteria, including the categorization detailed by Reinke and Zaretskaya. This study demonstrates that all the mentioned tools provide external integration capabilities. However, only a select few offer internal integration.

The interviews conducted as part of this study aimed to gather data on the integration of machine translation technology into translation workflows. Through these interviews, we sought to gain insights into the ways in which translation professionals are currently using machine translation technology, as well as their perceptions of the benefits and challenges associated with this technology.

3 Methodology

3.1 Data Collection

We conducted semi-structured interviews with each participant via video conferencing. The interviews were designed to explore participants' perceptions and experiences of working with MT in the context of medical translation. The interview guide covered questions around topics such as how participants use MT in their work, their opinions on the quality of MT output for medical texts, and their concerns about the impact of MT on the translation industry.

The interviews were audio-recorded with the participants' permission, and transcribed verbatim.

3.2 Data Analysis

We conducted a thematic analysis of the interview transcripts by generating codes based on our research questions and interview guide to capture key themes and ideas. The data was then organized and analyzed using the NVivo software.

3.3 Participants

We recruited participants for the interviews through professional networks and online communities for translators. Our only inclusion criteria were that participants had to be freelance translators in the medical field, either as part of their regular workflow or on an occasional basis. Participants did not need to have prior experience with MT. We sought a diverse group of participants in terms of experience, and familiarity with MT.

The participants in the interviews were voluntary and unpaid.

Participants were all freelance translators in the medical field, with an average of 17 years of professional experience in translation, and 70% having more than 10 years of professional experience. This suggests that they may have entered the field before post-editing training were available, and we observed that none of the participants reported receiving formal post-editing training during their university education. However, 9 participants reported receiving some form of short training on post-editing, either from a translation agency or through a professional development program.

It is worth noting that their limited training in post-editing may be a potential limitation for interpreting our results. As training is an important factor in determining the acceptability and efficacy of new technologies[9][4][7], the lack of training may have affected participants' attitudes towards and use of MT.

Additionally, our limited sample of 21 participants may not be fully representative of the large population of freelance medical translators, as those who have received more extensive training in post-editing may have different experiences and perspectives. Medical translation requires specialized knowledge and expertise in the medical field, as well as a deep understanding of the nuances of the target language. Therefore, it is worth noting that the strong professional experience of our participants may also be explained by the specialized nature of medical translation.

One limitation of our study is the small sample size, which is primarily focused on French-speaking translators, thus creating a narrow scope. The limited number of participants may restrict the generalizability of our findings to a broader population. Additionally, it is important to note that out of the total participants, 20 translators had English to French as their language pair, and only one translator had French to English as their language pair. This further emphasizes the imbalance in the sample and raises concerns about the representativeness of the data. Therefore, it is crucial to acknowledge that a more

diverse and representative sample, encompassing translators from various language combinations and geographical regions, would provide a more comprehensive understanding of the topic under investigation.

4 Findings

In this section, we report on the main outcomes of our qualitative investigation into the use of MT by freelance medical translators. Drawing on interviews with 21 professionals, we present the themes that emerged from our analysis and explore the motivations for the use of this technology, as well as the benefits and drawbacks professionals have experienced.

4.1 Acceptance of MT

Based on our analysis of the interview data, we divided our participants into two categories:

- those who accept MT,
- and those who refuse it.

The majority of our participants (17 out of 21) reported accepting MT to some extent, and only four participants expressed complete refusal of MT.

Participants Refusing MT Within our interview sample, the number of translators who refused to use MT was minimal, with only four participants expressing such sentiments. However, several trends can be observed in the nature of their responses, including the perceived limitation of creativity involved in the use of MT. Additionally, many of these participants cited at least one negative past experience with MT, mentioning instances of poor translation quality or an overall negative experience with an agency due to low rates or projects being sent for human translation revision despite suspicion of MT use. These negative experiences may contribute to the reluctance of some translators to adopt MT as a standard tool in their work, despite its potential advantages.

Participants Accepting MT Among those who accepted MT, we observed two distinct types of usage: passive and active. Passive usage refers to post-editing performed for translation agencies, without direct manipulation of the MT engines by the translators themselves. This type of usage is more common among those who work exclusively or primarily for translation agencies.

Active usage, on the other hand, refers to the autonomous use of MT engines by the translators themselves. This type of usage is more common among those who work for direct clients. Our findings suggest that active usage requires a different set of skills and knowledge regarding MT than passive usage, as it involves a more direct interaction with the MT engine.

Table 1. Active or Passive Usage of MT by Participants

Translator ID	PE for Translation Agencies	PE for Direct Clients	Active Usage
Trad01	Passive	-	-
Trad02	Passive & Active	Active	Yes
Trad03	-	-	-
Trad04	Passive	-	-
Trad05	Passive	-	-
Trad06	Passive & Active	Active	Yes
Trad07	-	Active	Yes
Trad08	Passive & Active	Active	Yes
Trad09	Passive & Active	Active	Yes
Trad10	Passive & Active	-	Yes
Trad12	Active	-	Yes
Trad14	Passive & Active	Active	Yes
Trad15	Passive	-	-
Trad17	Passive & Active	Active	Yes
Trad18	Passive & Active	-	Yes
Trad19	Active	Active	Yes
Trad20	Passive	-	-

The data comparing passive and active MT usage among our participants is summarized in Table 1.

The interviewees cited several reasons for actively using MT. One significant advantage is the ability to meet the demands of a competitive market and tight deadlines while increasing their productivity. Additionally, participants noted ergonomic reasons, such as reducing typing on the keyboard, as a benefit of using MT. One of reasons for using MT is also as an aid to translation through the proposals offered by the technology. This not only enables them to meet tight deadlines but also to accept more projects and increase their overall productivity. By leveraging the capabilities of MT, translators can streamline their work processes, thereby improving their efficiency and overall output.

Interestingly, some of our participants reported an active use of MT on human translation projects sent by translation agencies. Given the fact that some agencies explicitly prohibit MT use for human translation projects, we wanted to explore this data more in depth.

Among the 10 participants who reported having an active usage of MT on projects sent by translation agencies, 3 of them indicated that they used MT even though the contracts or instructions sent by the agencies explicitly prohibited it. They justified this action by stating that they believed that the quality of their post-editing was similar to what they could produce on their own without MT, and that they carefully post-edited the output to ensure that it met their own high standards of quality and accuracy. Additionally, they noted that they were using a paid version of MT software that guaranteed the confidentiality of the data.

4.2 Integration of MT and TM: Tools

According to the participants in the study, there are several ways in which MT is actively used. The majority of them explained that this integration takes place within their environment, either through batch processing carried out by a translation agency or through interactive processing when using MT actively and independently.

Only 2 translators explained that they use MT directly in their web browser, typically in cases where they have not invested financially in a dedicated MT tool or when their tool does not offer free integration, or on an *ad hoc* basis to assist with a particularly challenging translation. Overall, these different approaches to using MT highlight the flexibility and versatility of this technology for translation professionals.

Table 2 displays the tools used by translators reporting an active usage of MT in their CAT environment. Some translators reported using adaptive MT

Table 2. MT Integration into CAT Tools

Translator ID	CAT Tool	MT Engine
Trad02	MemoQ	ModernMT
Trad06	Trados	Cloud Studio
Trad07	MemoQ	ModernMT
Trad08	Trados	DeepL
Trad09	Café Tran	DeepL
Trad10	Trados	DeepL
Trad12	Memsources	DeepL
Trad14	Trados	DeepL
Trad19	MemoQ	ModernMT

engines, such as ModernMT, which incorporates post-editing feedback to improve the quality and accuracy of output over time. Several of these translators reported a perceived improvement in the quality of MT output when using adaptive MT engines, highlighting the potential benefits of using more advanced MT technology in translation workflows.

Several translators expressed concerns over the lack of customization options for punctuation settings in the MT engines they used, which led them to develop workarounds such as conducting a global search and replace of punctuation marks at the end of the post-editing process. This is an important ergonomic issue that highlights the need for further development in MT technology. By providing more customizable options for punctuation settings, MT developers could improve the usability and efficiency of their products and increase the satisfaction of post-editing professionals. The need for punctuation and customization with machine translation is particularly relevant in our sample, as all participants had at least French and English as language combinations, and these two languages present differences in punctuation. For instance, during the interviews participants expressed a need for the ability to customize their preferred apostrophe

or quotation mark styles. They highlighted that having the option to configure these punctuation marks according to the preferences of the specific project or client would save them the time and effort of performing a search-and-replace operation at the end of their post-editing tasks.

This feedback underscores the importance of providing translators with flexible settings that align with their specific language and formatting requirements, ultimately enhancing their efficiency and workflow. As such, addressing this issue represents an important opportunity for technological improvement in the translation industry.

Moreover, many translators reported encountering frequent inconsistencies in terminology and phraseology during post-editing, which they found frustrating and time-consuming. These inconsistencies can lead to a shortfall in the overall quality of the translated content, and can be particularly challenging in medical translation. However, translators using adaptive MT engines, which incorporate post-editing feedback to improve the quality and accuracy of output over time, reported less inconsistencies. This suggests that the use of adaptive MT engines may represent an effective strategy for reducing the number of inconsistencies encountered during post-editing, and improving the overall quality and consistency of medical translated content.

4.3 Appreciation of PE

Among the participants who accept post-editing, either actively or passively, there is a wide range of attitudes and preferences towards this task. Some participants reported that they only engaged in post-editing because of market pressures and a reduced demand for human translation, and do not particularly enjoy the task. On the other hand, others have fully embraced post-editing and prefer it to purely human translation.

On average, among those translation professionals who reported accepting post-editing tasks, machine translation technology was present in approximately 55% of their total revenue. However, it is important to note that this percentage varies widely across individual translators, with some relying heavily on this technology for the majority of their work. Three participants in our study reported that MT was now involved in 100% of their total revenue, which means that they accept post-editing projects from agencies, but also use machine translation technology actively on all other projects, including traditional translation projects from agencies or direct clients. This suggests that for some translation professionals, post-editing may not only be a means of satisfying client demand for faster and cheaper translation, but could also show a high degree of satisfaction with the technology and a complete integration of machine translation into their workflow.

5 Conclusion

In conclusion, our survey has revealed a wide range of attitudes and practices when it comes to using MT and post-editing among freelance translators in the

medical field. While some respondents expressed a reluctance to the use of MT due to concerns about accuracy and quality or to the limitation of creativity when using MT, others reported using it actively or passively in different forms. These findings highlight the importance of recognizing the diverse needs and preferences of translation professionals when it comes to MT and post-editing, and the need for continued research and development in this field.

It is important to note that our sample size was relatively small and limited to a specific domain of translation, namely medical translation, and that our sample consisted of individuals with limited training in MT and post-editing, and as such, our findings should be considered preliminary. Future research should explore the attitudes and practices of a more diverse range of translation professionals in order to gain a more comprehensive understanding of the challenges and opportunities associated with MT and post-editing in different fields. Therefore, our findings may not be generalizable to other domains or to the broader translation profession. Moreover, medical translation is a highly specialized field that requires a particular set of skills and knowledge, which may not be representative of other areas of translation. In order to obtain a more comprehensive understanding of the attitudes and practices of translation professionals with respect to MT and post-editing in different fields, future research should aim to replicate our study with larger and more diverse samples across different domains of translation.

Our interviews also entail biases of subjectivity as participants report their own feelings and experiences. It is important to acknowledge that individual perspectives and personal interpretations may influence the information shared during the interviews. Each participant brings their unique background, beliefs, and biases, which can shape their responses. While efforts were made to maintain objectivity and encourage participants to provide objective feedback, it is crucial to consider the potential impact of subjectivity on the data collected.

Nonetheless, our survey suggests that MT and post-editing are being used in a variety of ways by translation professionals, and certain cases raise questions of professional ethics and underscore the need for ongoing education and training in best practices for data privacy and protection when MT is used on projects sent by translation agencies. By developing a more nuanced understanding of the use of MT and post-editing in translation, we can better support the needs of professionals in this field and ensure that these technologies are being used effectively and responsibly. Future research should continue to explore the potential benefits and challenges of using adaptive MT engines in translation workflows, and develop best practices for integrating this technology into existing translation processes. However, the potential benefits of adaptive machine translation need to be studied in the context of freelance translators, considering the diverse range of content they work with, even in a specialized field, and the variety of clients they serve. Each project comes with its specific terminology, style, and specific requirements. Therefore, evaluating the effectiveness of adaptive machine translation in this context becomes crucial. Freelancers also cater to a diverse clientele, including individuals, small businesses, and multinational corporations,

each with distinct expectations and quality standards. Understanding how adaptive machine translation can align with the needs and preferences of freelance translators and their clients is essential for assessing its potential impact on their workflows, productivity, and the overall quality of translations delivered.

Building on the findings of our survey, we plan to further explore the use of MT and post-editing in the context of medical translation. More specifically, we aim to conduct a quality evaluation of post-editing in an active use setting on medical documents. By doing so, we hope to gain a deeper understanding of the practical implications and challenges for freelance translators using MT in a domain where accuracy and compliance are of paramount importance. This research could provide valuable insights into how MT and post-editing can be used most effectively and responsibly in a medical translation context, and could inform the development of new tools and strategies for improving the quality and efficiency of medical translation.

References

1. 2018 EUROPEAN Language survey results announced - EUATC (Apr 2019), <https://euatc.org/industry-surveys/2018-european-language-survey-results-announced/>, section: INDUSTRY SURVEYS
2. Forcada, M.L.: Making sense of neural machine translation. *Translation spaces* **6**(2), 291–309 (2017), ISBN: 2211-3711 Publisher: John Benjamins Publishing Company Amsterdam/Philadelphia
3. Kanavos, P., Kartsaklis, D.: Integrating machine translation with translation memory: A practical approach. In: Proceedings of 2nd joint EM+/CNGL workshop “Bringing MT to the user: Research on integrating MT in the translation industry” (JEC ’10). pp. 11–20. Denver, USA (Nov 2010)
4. O’Brien, S.: Teaching post-editing: a proposal for course content. In: Proceedings of the 6th EAMT Workshop: Teaching machine translation. pp. 99–106. Manchester, UK (2002)
5. Quintana, R.C., Castilho, S.: A review of the Integration of Machine Translation in CAT tools. In: Proceedings of the new trends in translation and technology conference - NeTTT 2022 (Mar 2023)
6. Reinke, U.: State of the art in translation memory technology. *Translation Memory Technology, Translation: Computation, Corpora, Cognition* **3**(1), 27–48 (2013)
7. Vasconcellos, M., Leon, M.: Spanam and Engspan: Machine Translation at the Pan American Health Organization. *Computational Linguistics* **11**(2-3), 122–136 (1985)
8. Vidrequin, M.: Assessing Quality and Use of MT by Professional Freelance Translators in the Medical Field. In: Proceedings of the new trends in translation and technology conference - NeTTT 2022 (Mar 2023)
9. Wagner, E.: Post-editing SYSTRAN, a challenge for commission translators. *Terminologie et traduction* **3**, 1–7 (1985)
10. Zaretskaya, A., Corpas Pastor, G., Seghiri, M.: Integration of Machine Translation in CAT Tools: State of the Art, Evaluation and User Attitudes. Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry **8**, 76–88 (Jan 2015)

Workbench for Post-editing of Translations from English and Hindi to Dravidian Languages

Sobha Lalitha Devi, Pattabhi R K Rao and Vijay Sundar Ram

AU-KBC Research Centre, MIT Campus of Anna University, Chennai-600044

sobha@au-kbc.org, pattabhi@au-kbc.org, sundar@au-kbc.org

Abstract. In this paper we discuss about the translator’s workbench which provides the translator with various types of post-editing tools to facilitate the speeding up of the translation process. Post-editing is the process of correcting/ changing the machine generated translation to a syntactically and semantically correct translation. This work bench is specifically customized to handle the correction of translation from English and Hindi to Tamil and Malayalam from three machine translation systems, the Google unpaid version, Google paid version and Sampark System (Indian language to Indian Language MT system, Govt. of India). The tools in the workbench include the domain dictionaries, technical term translation correction, Grammatical correction modules where verbs are checked and corrected, Appropriate word selector with morph generator and interactive translation prediction where the users edits are automatically stored and predict if there is a similar error is encountered. We conducted a field trial of our post editor and it is found that it reduced the time taken to produce the final translation. This paper reports work in progress.

Keywords: TRANSLATORS WORKBENCH, POST-EDITING, DRAVIDIAN LANGUAGES

1 Introduction

With improvement of machine-translation (MT) technology and the demand of translation industry, post-editing has become an important production process in translation. Post-editors check and correct MT output to improve the translation quality and applicability. Translation post editing and revising are the two concepts in editing the translation output from any machine translation (MT) output and translation revision is considered as a “function of professional translators in which they find features of the draft translation that fall short of what is acceptable, as determined by some concept of quality, and make any needed corrections and improvements” [1], and it has an important role in ensuring the quality of the translation. According to ISO 17100, post-editing the other concept is “editing and correcting machine-translation output”.

A Translator's workbench provides the user with a set of computer-based tools which help the post-editors for speeding up the translation process. Post-editing is the process where humans identify the errors in machine-generated translation to achieve an acceptable final product. Various new interactive tools are developed with new features aiming to assist the post-editors which are included in workbenches. It is found that the translators are able to use the workbench without much training given [2,3].

With successful use of Moses, Statistical machine translation, Computer Aided Translation (CAT) tools integrated with Machine translation system and post-editing the output got prominence in translation industry. [4] presented PET: a Tool for Post-editing and Assessing Machine Translation, which mainly focussed on facilitating the post-editing of translations from any MT to reach publishable quality and also to study the sentence-level information from the post-editing process, e.g.: post-editing time and detailed keystroke statistics. [5] presented the details of post-editing work using AutoDesk tool in a translation industry, where they handled translation in 13 languages. [6] presented an open source CAT tool, MateCat tool, which was integrated with translation memories, terminology bases, concordancers, and machine translation system. [7] presented INMT: Interactive neural machine translation prediction, to assist human translators with on-the-fly hints and suggestions to achieve the end-to-end translation process faster, more efficient, and creates high-quality translations.

Multi-model post editing tool (MMPE) combining traditional input modes with pen, touch, and speech modalities for PE of MT was presented by [8]. The results of this evaluation with professional translators showed that pen and touch interaction were suitable for deletion and reordering tasks, while these instruments were not used for longer correction. IntelliCAT, an interactive translation interface with neural models to streamline the post-editing process was presented by [9]. They have shown a 52.9% speedup in translation time compared to translating from scratch.

UDAAN, is an open-source post-editing tool built with experiments from English to Hindi translation task. It has machine translation system along with post-editing tools. It provides 100 in-domain dictionaries for aiding the post-editing task [10]

Though there are post-editing tools available in European and English, there are very less attempts for Indian languages. And oddly any tool specifically catering to Dravidian languages such as Malayalam, Tamil, Telugu, Kannada, which are morphologically rich and requires special processing with post-editing the translation.

In this paper we describe our on-going work of developing a translator's workbench for MT output. A need of a translator's workbench was felt while we were working on translating undergraduate and post graduate lectures from University Grant Commission of India, the courses taught for Indian college students, (<https://pmevidya.education.gov.in/swayam-portal.html>) under "Swayam Platform". The courses were in 7 domains, Law, Environment, Mathematics, Chemistry, Biology, Computer Science and Public Administration. Since we had to deal with different domains and no MT system is customized for separate domains, the output we were getting from generic MT systems needed post editing at various levels. Purely manual post editing took enormous amount of time for the final product. There were 86 courses, a total 1300 hrs of lecture with a total of 3,50,000 sentences in English. The

paper is designed as follows, in the next section we discuss about the common errors from the three MT systems for English and Hindi to Dravidian Languages a) The Sampark system, b) Google free version and c) the Google paid version. In the third section we have given in details the architecture and the working of our workbench and finally the conclusion.

2 The Three Machine Translation Systems

The availability of automatic translation systems in most of the languages brought by the tech giants such as Google, Microsoft, Systran, Facebook etc has brought a paradigm shift in the translation industry. Though the translations by these automatic translation system needs correction, translators use these translation output as the base and perform post-editing of these translation to get proper translation. In this section, we give in detail the translation systems used in our workbench and discuss the errors in the MT translation output.

Workbench is designed to handle the translation output in the following language pairs, English-Hindi, English-Malayalam, English-Tamil, Hindi-Tamil, Hindi-Malayalam and Tamil-Malayalam. We have integrated three different translation systems with this workbench, namely, Sampark translation system, Google paid and free translation versions.

Sampark Translation System: Sampark translation system is built using Analysis-Transfer-Generate architecture using hybrid techniques. Modules were built using Machine Learning techniques with linguistic features and post-processed with rich set of linguistic rules. The Analysis part has the modules to perform a detailed analysis of the source sentence such as morphological analyzer, POS tagger, Chunker, Clause boundary identifier and Named Entity Recognizer. This is followed by the Transfer part, where the lexical, structural and syntactic transfers are done as required by target language using transfer grammar and lexical transfer engine. In the Generation part, target language sentence is generated using the lexical, structural and syntactic transfers performed in the transfer part. These systems were built under a funded project by Government of India. We use Hindi-Tamil, Tamil-Hindi, and Tamil-Malayalam translation systems built under this project for our workbench. They are available on http://www.tdil-dc.in/index.php?option=com_vertical&parentid=74&lang=en. We have included Sampark System to our workbench.

Google Translation: We have included both the paid and free translation services provided by Google, which are built using neural machine translation techniques, into our workbench. The free version has character limitation and limitations on number of usage. It fails very often. To overcome these issues, we have included the paid version of Google translation API to our workbench.

2.1 Errors from The Three Machine Translation :

We broadly classify the errors in the translation as simple and complex errors. Simple errors include spell errors, case transfer errors, copula errors and errors in selection of Domain /Technical terms, correct lexical item as per the context of the sentence, Translation of Named Entities. Complex errors include the sentence construction error and error in generation of complex verb phrases. We explain in detail each of the errors with example.

Case Transfer Error: As Indian languages are inflection and morphologically rich, the case marking with the nominals play an important role in semantic interpretation. There is no one to one mapping of case markers between languages as in locative case in Malayalam need not be locative case in Tamil. There are also one to many and many to one case mapping between Indian languages. ‘se’ “instrumental” in Hindi can be transferred into ‘aal’ instrumental, ‘ai’ accusative, ‘ilirunthu’ ablative case in Tamil. The context will determine which case should be taken while translating from Hindi to Tamil. The locative case in Malayalam changes to dative in Tamil. Consider the example. ML: avan chennaiyil (**il is the locative case**) poyi; TA: avan chennaikku(**kku is the dative case**) poonanan. (He went to Chennai.). Both NMT systems and Sampark systems introduce this errors in case transfer. Sampark system has specific linguistic module to handle case transfer in transfer grammar engine, so it performs better than Google translation engines.

Copula Generation Error: Hindi and Malayalam have copular construction and it is a necessary condition for the sentence to be syntactically correct, whereas it can be dropped in Tamil. In the translation from Tamil to Hindi and Malayalam copula generation error occurs. In Sampark translation engine, we have a copula generation engine and it is handled better in Sampark system output and these errors are more in Google translation. Copula errors often occur in Tamil to Malayalam and Tamil to Hindi translation output.

Lexical Selection: Correct selection of context specific lexical item is a challenge and this is not handled by any of the three translation engines. It is comparatively less in Google paid version, as it is trained with large data. The lexical selections are of three types: 1. Domain/Technical Term Selection, 2. Context depended term selection and 3. Named Entity selection. The Domain terms and context terms are not properly selected by all three systems but Named entity is handled by Google paid version and Sampark systems.

Another issue in term selection is regarding equations in Mathematics and Chemistry. They are not handled by all three systems.

Errors in Complex Construction: Translation of Multi-clause (sentences with multiple embedding) and long sentences using Google brought in error in the target language syntactic construction and require editing at a higher level. Sampark system on the other hand, handles these sentences better than Google as it analyses the source language at the syntactic level.

Verb Transfer: Sampark engine having word generation using linguistic information generates more precise verb forms than Google translation. In Google translation we get incorrect verb forms instead of the exact verb form. Both Malayalam and Tamil, being morphologically rich languages, have complex verb generation.

Unknown Word Handling: Unknown words are the potential challenge in Translation output. Sampark systems fail to handle the unknown words. It will just transliterate those words. It will not even generate the word with proper inflections. This affects the translation of the sentence. In the case of Google translation unknown words are handled using different techniques and presents it as either transliteration word or generate into another word using sub-word techniques. In both the cases, validating these words is required.

Degree of Translation Error: Sampark and Google systems produce perfect translation at the syntactic level for simple generic sentences whereas they bring in errors if the sentences are complex. The lexical errors are common in both the constructions. Google paid version is better than the free version, but both the engines fail in translation of complex sentences with multiple embedding.

3 Workbench and Its Architecture

The working of our translator's workbench is as follows. The user can upload the input data for translation in any of the format (doc, docx, pdf, txt) and can choose the target language. After choosing the target language the user can choose the translation engine among the three translations systems (Google translation (paid and free service), and Sampark translation) for English- Hindi, English-Tamil, English-Malayalam, Hindi-Tamil, Hindi-Malayalam and Tamil-Malayalam translations. The architecture of the workbench is presented in figure 1 (Fig is attached at the end of the paper after reference). The output in the target language from the translation systems and the input source language are shown to the Post editor. The editor can see the target sentence and mark it as correct or not correct by clicking yes or no. If it is yes it will change the colour to green and can move to the next sentence. If it is not correct then it will show all the tools available on either side of the tool box in a hierarchical order, while correcting at each hierarchy, the next level will be shown. The workbench itself will show certain level of errors through its intelligent Error identifier which is equipped with automatic NE identifier, Domain Term Identifier and Technical, mathematical names.

3.1 Translation Error Handling

The tool identifies errors in translation using a NER, Domain Identifier and Terminology Identifier. It also identifies partially translated sentences and certain improper translation and alerts the user. The partial translation is identified by comparing the number of characters and words in the input sentence and the translated output sentence. When there is drastic difference, then an alert message is given to translator

while he corrects that sentence. Consider the following example of English to Malayalam translation.

Input: *For instance, if individuals with capital 'T' are more successful in reproduction than the individuals with small 't', the frequency of the former will be higher.*

Translation Output:

Mal: ഉദാഹരണത്ത ന്,

(For example, capital(N)+ulla(ensp) individual(N)+plural+copula+conj)

Considering the above example, the input sentence has 29 words and 179 characters and translated sentence has 4 words and 49 characters. While comparing English and Malayalam we expect the number of words in Malayalam to be more than 40% of words in English. If the translation has less than 40% of words then it is flagged as partial translation. This value 40% is obtained by the empirical analysis of the English-Malayalam parallel data.

Improper Translation: Improper translation is identified using pattern analysis. Consider the following example (English to Malayalam translation).

Input: *The second website is www.theagricos.com.*

Translation Output: രണ്ടാമത്തെവെബ്സൈറ്റ് www. . കൂടെ .

The input sentence has an url and in the translation it should remain the same. But the part of the url is translated. This will be identified as improper translation and user will be alerted.

3.2 Domain Based Translation Correction

Workbench has automatic domain classifier and it will classify the domain of the input and choose the appropriate dictionaries for the target language. The tool will highlight the Technical and Domain Terms in source language and the editor can use that to correct the output. We have domain and technical term dictionary, mathematical dictionary, Chemical name dictionary and botanical name dictionary. Domain and technical term glossary for look up is included in the workbench and also has domain and technical term dictionaries used for identifying the terms and highlighting it.

Since Malayalam and Tamil are morphological rich languages, while replacing the lexical items, suffix attached to the lexical items has to be handled. Consider the following example, which has English to Tamil translation.

Input: *In addition to the nuclear genome plants also contain genes in the organisms like chloroplast or plastids and mitochondria.*

Translation Output:

ஆலைகளுக்க, க,

ன.

Corrected Translation:

ரக,

ன.

(anukkaru marapaNu(N) thaavarangkaLukku kututhalaaka kuloro-
 piLaast nuclear(N) genome(N) plant(N)+PL+DAT additional(N)+ADV chlo-
 roplast allawu piLaastidkaL maittokaaNriyaa uyirinangkaLilum or(CONJ)
 plastid(N)+PL mitochondria(N) organism(N)+PL+INC marapaNukkaL ullana
 gene(N)+PL is(V)+past.)

In the above example, the lexical item ‘plants’, is translated as ‘ஆலகைகளுக்க’ (for industries), which means ‘for the industries’, but the sentence is from biological domain and the lexical substitution has to be ‘பிளாஸ்டிட்காலை’; (for plant). In this example ‘ஆலகைகளுக்க’ has plural marker ‘கள்’(kaL) and dative case marker ‘க்கு’(kku) suffixed to the root noun ‘ஆலை’. While replacing ‘ஆலை’ with ‘தாவரம்’, the lexical item has to be generated for these two suffixes as ‘தாவரங்களுக்கு’. Here we accomplish by the identification of the suffixes and generating the new word with the required morphological suffixes by using a light weight morphological analyser and morphological generator built using finite state automata and paradigm dictionaries. In the present system we have 15 different domain dictionaries.

3.3 Spelling Correction and Alterations

After the translation corrections are done, we ensure the spelling correction in the translation. We have in-house built spell checkers for Malayalam and Tamil using n-gram and finite state automata techniques. If the spell checker identifies the error word, it will be highlighted and the possible suggestion words will be provided to the user as a suggestion list. Here we also take care of the named entities, as the spelling of the NEs has to be consistent throughout the document. Using in-house developed Named Entity Recognizer (NER), we identify the NEs in the source sentences and spelling for those NEs in the translated sentences are maintained consistently.

For Tamil text, we have sandhi-validation module. In Tamil we have to maintain the lengthening of sound between the words by marking the end of the previous words with vowel. Consider the following example, பள்ளிப் படிப்பு (school education) Between the two words ‘ப்’ a vowel is inserted as the sandhi (morphotactic rule). These sandhi validations are done using a set of linguistic rules.

Evaluation and Conclusion

We have evaluated our workbench using the data from Swayam courses. The total number of English sentences we have used through the workbench is 3,50,000 sentences, Number of simple and complex sentences are 1,15,500 and 2,34,500 respectively. Other statistics are given in the table 1.

Table 1. Sentence Wrongly Translated by Systems

S.No	Details	Google paid (English to Tamil)	Count in Sampark (Tamil to Malayalam)
1	Completely Wrong Translation in Mathematical Domain	97500	87500
2	Completely Wrong Translation in other Domain	15000	11450
3	Number of Sentences where NE/Domain Term has to be corrected	8000	9500
4	Number of Sentence where Mathematical Equations has to be corrected	114000	10900

We have employed 20 Tamil and 20 Malayalam editors to correct the machine output. The time taken for correction by editors is as given in Table 2

Table 2. Time Taken by Editors

S.No	Sentence Type	Time Taken
1	Simple Sentences	30 sec to 1.5 mins
2	Complex Sentences	3 to 5 mins
3	Sentences with complex Mathematical Equations	5 to 15 mins

We have discussed about our translators workbench specifically customized to handle the correction of translation from English and Hindi to Malayalam and Tamil using three translation systems, namely, Sampark system, Google paid version and Google free version translation service. We have described the errors in the translation output of these systems. Domain dictionaries, technical term translation correction and grammatical correction modules are included in the workbench. These features help in speeding up the post-editing task.

References

1. Koponen, M. Brian Mossop, Isabelle S. Robert, Giovanna Scocchera. Translation Revision and Post-editing: Industry Practice and Cognitive Processes. Routledge. <https://doi.org/10.4324/9781003096962>. (2020)
2. Maarit Koponen, Wilker Aziz, Luciana Ramos and Lucia Specia. Post-editing Time as a Measure of Cognitive Effort. In Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice, pages 11–20. San Diego, California, (2012).
3. Wilker Aziz, Maarit Koponen and Lucia Specia. Sub-sentence Level Analysis of Machine Translation Post-editing Effort. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia (eds). Post-editing of Machine Translation: Processes and Application, pages 170–199. Newcastle upon Tyne: Cambridge Scholars Publishing, (2014)

4. Aziz, Wilker and Castilho, Sheila and Specia, Lucia. "PET: a Tool for Post-editing and Assessing Machine Translation", In: Proceedings of the Eighth International Conference on Language Resources and Evaluation, Istanbul, Turkey (2012)
5. Ventsislav Zhechev, "Machine Translation Infrastructure and Post-editing Performance at Autodesk" In book: Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP '12) Editors: Sharon O'Brien, Michel Simard, Lucia Specia, (2012)
6. Marcello Federico, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, et al. The MateCat tool. In COLING (Demos). 129–132. (2014)
7. Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. INMT: Interactive neural machine translation prediction. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. 103–108. (2019)
8. Nico Herbig, Tim Düwel, Santanu Pal, Kalliopi Meladaki, Mahsa Monshizadeh, Antonio Krüger, and Josef van Genabith. MMPE: A multi-modal interface for post-editing machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 1691–1702. (2020)
9. Dongjun Lee, Junhyeong Ahn, Heesoo Park, and Jaemin Jo. IntelliCAT: Intelligent Machine Translation Post-Editing with Quality Estimation and Translation Suggestion. arXiv preprint arXiv:2105.12172 (2021).
10. Ayush Maheshwari and Ajay Ravindran and Venkatapathy Subramanian and Ganesh Ramakrishnan ., "UDAAN: Machine Learning based Post-Editing tool for Document Translation, In: Computation and Language. (2022)

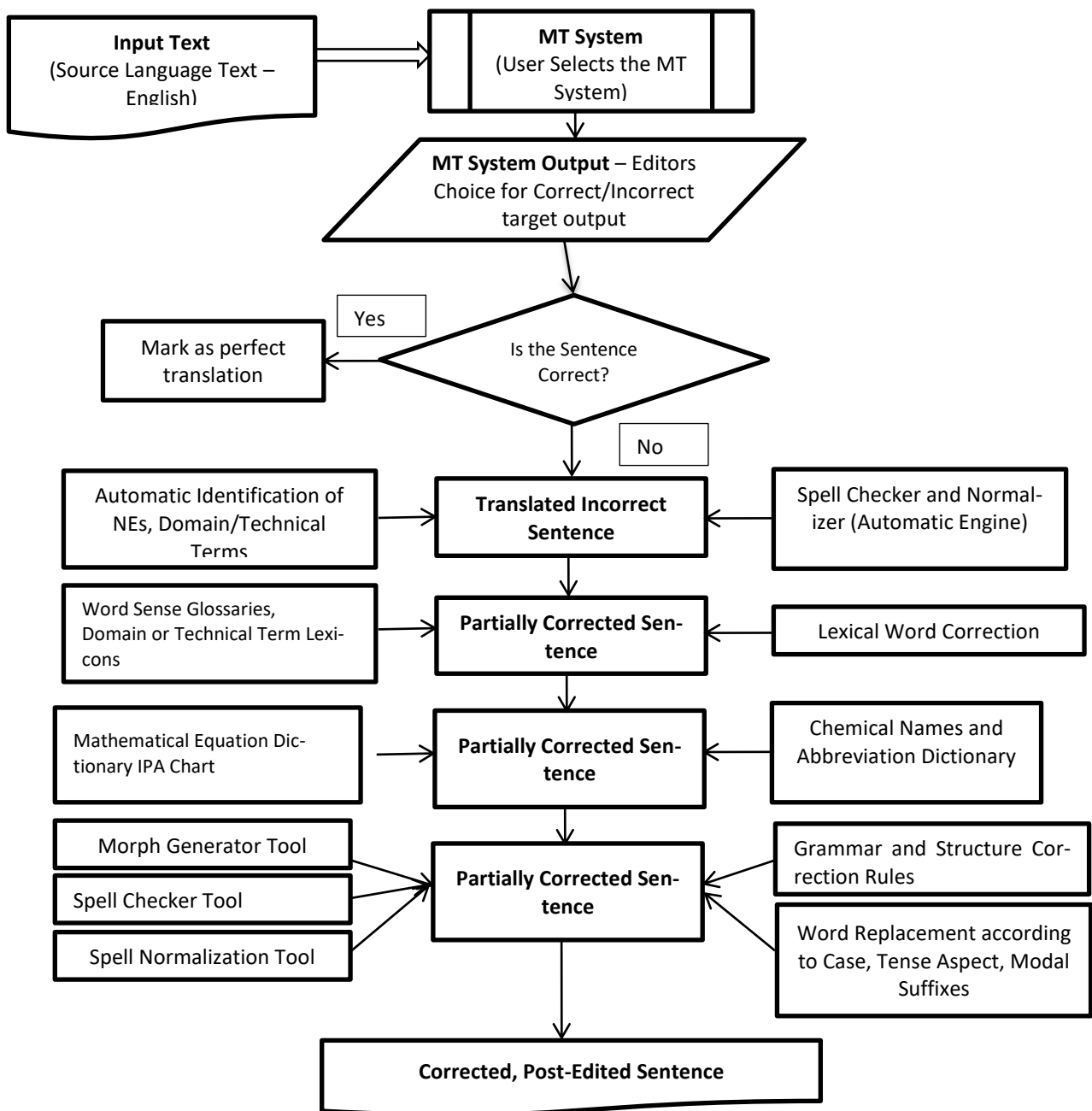


Fig. 1. Architecture of the Workbench

Processing English Verb Phrase Ellipsis for Conversational English-Hindi Machine Translation

Aniruddha Prashant Deshpande¹[0009–0005–4072–3718] and
Dipti Misra Sharma¹

International Institute of Information Technology - Hyderabad
Hyderabad, Telangana 500032, India
{aniruddha.d,dipti}@research.iiit.ac.in
<https://ltrc.iiit.ac.in>

Abstract. In this paper, we try to tackle the problem of erroneous English-Hindi machine translation (MT) outputs due to the presence of the Verb Phrase Ellipsis (VPE) in English. The phenomenon of VPE is prominent in spoken English, and the antecedent to the ellipsis can come from previous sentences in a conversation as well. MT systems translate sentences as a whole and ignore the contextual information from the previous sentences. For these two reasons, spoken English-Hindi translations suffer. We approached this problem by manually annotating 1200 two-person conversations that contain VPE and by studying how their resolution affects the translation qualities. Using our studies, we designed a rule-based system for the detection and resolution of VPE in English with the goal of improving their subsequent Hindi translation qualities. Our rule-based system is capable of the following: 1) Detection of VPE, 2) Resolution of Elided Head verb, 3) Resolution of Elided Head verb's children, 4) Resolution of non-verbal predicates of a copula or a 'be' main verb, 5) Modifying original sentence in the conversation with the resolved verb phrase. We also tested the system's performance on VPE datasets outside of our annotated data. In this paper, we present our annotated corpus on conversational English VPE, our rule-based system to tackle VPE in the context of improving English-Hindi MT, the observations made as we designed this rule-based system and the performance-related observations of our system.

Keywords: Verb Phrase Ellipsis · English-Hindi Translation · Rule-based system

1 Introduction

Ellipsis refers to a linguistic occurrence where certain syntactic elements are left out but can be inferred from the surrounding context. In general, for a sentence to be complete, it should contain a verbal constituent. However, sometimes sentences can be found that do not contain an explicit verb form in order to avoid

redundancy; yet they are intuitively complete. [1] This specific form of ellipsis that involves the omission of verb phrases is known as Verb Phrase Ellipsis (VPE). It is observed that a Verb Phrase Ellipsis (VPE) is comprised of two components: a **licensor**, which is usually an auxiliary or a modal verb that signifies the occurrence of a VPE, and an **antecedent**, which refers to the verb phrase that the omitted element corresponds to. [2, 3]. This is illustrated in the English sentences within Example 1 & 2 below.

State-of-the-art MT systems often fail to interpret and translate sentences with VPE. This error is further propagated when the antecedent to the VPE is present in one of the previous sentences that adds context to the current sentence that is being translated. This is so because these MT systems translate sentences as a whole, and the context provided by previous sentences is forgotten. Note that the antecedent for VPE may be present in the same sentence (Example 1) where VPE is detected and also in one of the previous sentences (Example 2). In our paper, we specifically study how the presence of VPE in English affects the quality of English-Hindi MT systems in a conversational context. The following two examples illustrate the erroneous Hindi translations due to the presence of VPE in English. (In both examples, the licensor is depicted in italics with a subscript L and its corresponding antecedent is depicted in bold. In Example 2, A & B denote the speakers in a conversation):

Example 1: Erroneous Hindi MT output of English sentences where VPE and its antecedent is present in the same sentence:

1.1) English: Maybe some animals **eat the seed**, but humans *don't_L*.

1.2) Hindi: ho sakta hai ki kuch jaanvar beej khaate hon, lekin manushya nahi.

Example 2: Erroneous Hindi MT output of conversational English sentences where VPE's antecedent is present in a previous sentence:

2.1) English: A - By the way, do you **know the price of the gas** now? B - I'm afraid I *don't_L*.

2.2) Hindi: A - waise, kya aap abhi gas kee keemat jaante hain? B - Mujhe darr hai main nahi karta hun.

Observing such errors motivated us to tackle English VPE in a conversational context, the resolution of which would assist English-Hindi MT systems. For our study, we manually annotated 1200 two-person conversations that contain VPE from the DailyDialog dataset [4]. Inspired by N. Bafna et al. [5], we developed a rule-based system based on observations made on our data that 1) Detects VPE, 2) Resolves Elided Head verb, 3) Resolves Elided Head verb's children, 4) Resolves non-verbal predicates of copula or "be" main verb, 5) Modifies original sentence in the conversation with the resolved verb phrase. The annotated data and the rule-based system are available here¹. Finally, we studied the performance of our rule-based system on our annotated data; tested the system's performance on 50 new conversations (outside our annotated corpus) and even on 50 new instances taken from the annotated data provided by Bos and Spe-

¹ <https://github.com/aniruddhapdeshpande99/VPE-Processing-EN-HI>

nader [2] to allow us to discuss its advantages and shortcomings as a part of our study.

2 Related Work

Several previous studies have tackled the issue of resolving antecedent heads for VPE in English. Hardt[6, 7] conducted the initial investigation into the detection of VPE and the identification of antecedents through computational and algorithmic methods. Hardt’s approach [6] involved using linguistically motivated rules that were heuristic in nature to resolve VPE. Nielsen [8] introduced the initial end-to-end system that solves VPE from unprocessed text input. He details various heuristic and learning-based techniques for identifying targets and antecedents. Nielsen [8] also provides high-quality annotated data on VPE. This data was also expanded upon by Bos and Spenader [2] and has been a popular dataset to study the phenomenon.

ViPER [9] makes use of string-based rules to detect and resolve VPE whose antecedent occurs in the same sentence. The authors of the paper designed their system first to identify only the cases that can be easily resolved and utilizes rules that work with high precision to detect and resolve VPE. Their strategies involve phrasal pattern matching, looking for simple structurally parallel contexts, and finally, looking for the modal verb that is semantically paired with the VPE licenser. Liu et al. [10] explored this problem with the use of joint learning. Kenyon-Dean et al. [11] approached the problem with the use of supervised discriminative machine-learning techniques and made use of the Margin-Infused-Relaxed Algorithm for resolution. R. Aralikatte et al.’s approach [12] to VPE involved using architectures developed for question answering.

Research on VPE affecting English-Hindi MT has been conducted by N. Bafna et al. [5], wherein they proposed a solution that made use of VVPE in Hindi. VPE in Hindi works differently compared to VPE in English. Hindi exhibits the phenomenon of the verb-stranding verb phrase ellipsis (VVPE) [13]. In VVPE, the verb moves out of the verb phrase to a higher head, and then the verb phrase is elided (including all remaining VP-internal material), stranding the verb itself. They proposed resolving only the head verb antecedent at the site of the ellipsis, essentially stranding it without its objects and other arguments, is sufficient to improve MT results on the modified sentences. They developed a rule-based system with the idea of transferring English VPE to Hindi VVPE on single-sentence data from the WSJ and the BNC corpus [14]. The presence of VPE in English affects the quality of MT outputs for languages other than Hindi as well. P. Khullar’s study [18] illustrates the effects of VPE when translating from English to Telugu. Beyond Indian languages, the presence of VPE in English also affects translation quality when translating into Persian [1] and Russian [19].

3 Our Approach

We decided to approach our problem with a rule-based system. Our reasoning for doing so stems from the lack of data on conversational English VPE. VPE resolution is a natural language generation task and would require a lot more data in order to train a deep learning-based model. Therefore, we decided that a rule-based system would be more apt for our problem task. We expand upon the system designed by N. Bafna et al. [5] to handle cases in a conversational context where the antecedent can also come from previous sentences. Our rule-based system is implemented in Python and takes in individual conversational texts as its input. The conversations are then parsed using dependency parsers. We then make use of conditional statements on the parsers' outputs to apply the rules to detect and resolve VPE within the conversation. The rules were designed after studying the Hindi translations of entire conversations containing VPE. We used Google Translate² and HimangY Translate³ to retrieve Hindi translations. The rule-based system is explained in detail in **Subsections 3.2 and 3.3**.

3.1 Data Creation And Annotation

Previously conducted research on the task of VPE usually utilizes annotations created by Nielsen [8] that take samples from the Brown, the WSJ and the BNC corpus [14]. They do not take into consideration pure dialogues or conversations. We wanted to study how resolving English VPE in conversations affects their subsequent Hindi translations.

Due to sparsity in annotated VPE data for conversational English, we decided to make use of the DailyDialog dataset [4] for our VPE annotations. The dataset consists of 13,118 two-person conversations, and on average, there are around 8 speaker turns per dialogue with around 15 tokens per turn. Within these, 1200 conversations with VPE were manually annotated, and we found 1658 total VPE instances. Since most instances of VPE are triggered by auxiliary verbs and the infinitive marker "*to*", we demarcated these explicitly in our annotations. The main five categories are 1) *be* licensors, 2) *do* licensors, 3) *have* licensors, 4) modal auxiliary licensors, 5) infinitival *to* licensors. However, we also noticed cases where VPE is triggered by non-auxiliary verbs, which were annotated separately. The dataset also contains instances of exophoric cases of VPE, i.e., the antecedent to the VPE comes from the context that is not present within the conversation. We also considered VPE in tag questions and ellipsis of non-verbal predicates within our dataset. Our annotation schema is summarized in Table 1 in the Appendix section. We divided our data into a group of 800 and 400 conversations for training and testing, respectively. By training here, we mean that we used our training set to study and develop our rules for both detection and resolution tasks.

² <https://translate.google.com>

³ <https://ssmt.iiit.ac.in/translate>

3.2 VPE Detection

Since we need outside context to resolve exophoric cases of VPE, we limited ourselves to detecting only the endophoric cases of VPE in our rule-based system. As observed by Bos and Spenader [2], VPE is most often triggered by auxiliary verbs and by the infinitive marker "to". We decided not to design rules for non-auxiliary licensor detection for the following reason. Within our training set, we only observed 39 instances of non-auxiliary licensors out of a total of 1114 endophoric VPE cases. This also matched the observation made by Bos and Spenader [2]. From these 39 instances, only 6 cases produced erroneous translations. This meant that non-auxiliary licensors that led to erroneous translations only amounted to 0.538% of cases of VPE that were observed. We decided not to consider non-auxiliary licensors due to the negligible level of frequency of the erroneous translations that they led to. As observed by N. Bafna et. al [5], VPE instances triggered by Do-So anaphora need not be resolved as they do not impact Hindi MT outputs negatively, therefore, they are ignored by our detection sub-system.

We first assume that if we encounter a token that belongs to the following classes: 1) *be* verbs, 2) *do* verbs, 3) *have* verbs, 4) modal auxiliary verbs, 5) token *to*, then it is a licensor. We then apply our detection rules to this assumption to eliminate cases where these tokens do not act as licensors of VPE. Note that we do not need to look at previous sentences to decide whether a VPE licensor candidate is actually a licensor or not. We utilized both Stanford CoreNLP [15] and Spacy dependency parsers [16] to decide this. We only considered cases where both the parsers' outputs are in agreement to reduce the error propagation that may arise from incorrectly generated dependency parse trees.

Detection Rules That Apply To All 5 Categories A token from the aforementioned 5 categories is not a licensor if it acts as an auxiliary to another verb or if it has a direct object child. These tokens will not act as a licensor if they are a parent to a clausal complement child. Verbs belonging to the *be*, *do* and *have* categories can also behave as main verbs; in such cases, they will not act as licensors of VPE.

In spoken conversations, we also observe a high presence of affirmative and negative sentences. They could either be a response to an interrogative Yes/No question or could be acknowledging/rejecting a statement that occurs in a prior context. A speaker may also ask an interrogative Yes/No question addressing a statement that was uttered previously in the conversation. A definite pattern that we see is that these types of sentences always involve auxiliary verbs. In these situations, the absence of predicates to the auxiliary triggers VPE. This is also true for VPE triggered in tag questions. We also discuss detection rules that apply to separate categories below.

Detection Rules For *be* Licensor A *be* verb token is not a licensor if it acts as a copula. We also need to identify a special case wherein the *be* verb acts as

an adverbial modifier parent to an adverb and distinguish whether it acts as a licensor or not. In this scenario, if the *be* verb's adverb child takes the semantic role of ARG2 (i.e. benefactive, instrument, attribute, or end state), then it is not a licensor. We used AllenNLP's Semantic Role Labeler [17] to identify the semantic role labels of the adverb. Example 3 illustrates this special case:

Example 3: Special case of *be* verbs as an Adverbial modifier where the *be* verb is a licensor:

Sentence 3.1: A - Juliette *:is:* going to the party.

Sentence 3.2: B - Roger *::is::* <> too.

In the above example, the *be* verb, which is demarcated between two double colons in italics, acts as a licensor (*::VPE_Licensor::*) within Sentence 3.2. The position where its respective antecedent VP is to be resolved is demarcated by <> . In Sentence 3.1, there is another *be* verb, which is not a licensor, which is annotated between two single colons and is in italics (*:Non_Licensor:*). We continue using these notations throughout **Subsection 3.2** to distinguish between a licensor and a candidate that is not actually a licensor.

The *be* verb "*is*" in Sentence 3.1 is not a licensor as it acts as an auxiliary to a gerund verb "going". In Sentence 3.2, the *be* verb "*is*" has an adverb child "*too*". However, since the adverb child doesn't take the semantic role of ARG2, we need to identify its parent *be* verb "*is*" as a licensor of VPE.

Example 4: Special case of *be* verbs as an Adverbial modifier where the *be* verb is not a licensor:

Sentence 4.1: A - Where *:is:* John?

Sentence 4.2: B - He *:is:* ahead.

Now consider Example 4. In both Sentence 4.1 and Sentence 4.2, we see that there is a *be* verb "*is*". In Sentence 4.1, the *be* verb has an adverbial child "*Where*", and in Sentence 4.2, the *be* verb has an adverbial child "*ahead*". The only difference in these sentences, compared to Sentence 3.2, is that these adverbial children take the semantic role of ARG2. Therefore, the *be* verbs in Sentence 4.1 and Sentence 4.2 do not act as licensors.

Detection Rules For *have* Licensor Beyond the general rules, we noticed that if a *have* verb has a nominal parent with a relative clause dependency relation between them, then it is not a licensor. In Example 5, the *have* verb "*had*" is not a licensor as it has a nominal parent "*day*" (underlined) linked by a relative clause dependency relation. We also see another *have* verb, "*ve*", in the following example, which is not a licensor as it is acting as an auxiliary.

Example 5: *have* verb which is a part of a nominal relative clause:

A - How was your day? B - Actually it was the most interesting day that I' *:ve:* *:had:* .

Detection Rules For Infinitival *to* Licensor We observed that a token *to* is not a licensor if it is not an open clausal complement child. It also does not

act as a licensor when it behaves as a preposition. Example 6 below illustrates the application of these rules.

Sample Run Of Our Detection Rules Here, using our rules above, we would like to show how they would identify VPE Licensors within a conversation. Consider the conversation from Example 6 below. The following example follows the same notation described in Subsection 4.2, and all the possible candidates are indexed with a number in the subscript. Candidates 1, 2, 5, 6, & 8 are not VPE licensors as they act as an auxiliary to another verb. Candidates 4 & 7 aren't licensors of ellipsis because they behave as a preposition. The remaining candidates do act as VPE licensors. Candidate 3, i.e. the token *to* doesn't act as a preposition, is not acting as an auxiliary, nor is it an open clausal complement child, hence making it a licensor. Candidates 9 and 10 are modal auxiliary candidates, and since they do not have a main verb parent, they are licensors of VPE. Candidate 11 is neither acting as a copula, nor is it an auxiliary to any main verb, nor is it a main verb to a predicate child. Therefore, it is a licensor too. Also, note that Sentence 6.4 exhibits the case of a non-auxiliary licensor of VPE. Here, the modal adverb "Certainly" acts as a licensor. Since we have kept such cases out of scope for our study, our detection system ignores it.

Example 6: VPE Detection Sample Example:

Sentence 6.1 A - Why *:don't:1* you sit down, darling?

Sentence 6.2 B - I *:don't:2* want *::to::3* <>.

Sentence 6.3 A - Well, come and talk *:to:4* me then.

Sentence 6.4 B - Certainly not.

Sentence 6.5 A - *:May:5* I turn on the radio, then?

Sentence 6.6 B - Turn on the radio.

Sentence 6.7 B - What for?

Sentence 6.8 A - So that we *:can:6* sit down together and listen *:to:7* some music.

Sentence 6.9 B - And who *:ll:8* cook dinner?

Sentence 6.10 B - *::Will::9* you <>?

Sentence 6.11 A - Ok, I *::will::10* <>.

Sentence 6.12 A - Cooking sounds easy.

Sentence 6.13 B - It most definitely *::is::11* n't <>.

3.3 VPE Resolution

For the resolution task, our system works to resolve both the elided main verb and its children. Contrary to the system designed by N. Bafna et al. [5], we decided to opt against only resolving the antecedent head verb. This is so because we observed that doing this doesn't necessarily lead to better translations for conversational data. Even if the translation is fluent, its adequacy with respect to the context provided by the previous sentences is lost. Example 7 illustrates this drawback. In the below example, just resolving the head antecedent verb "*walk*" won't suffice, as by leaving it stranded, we receive a translation that signifies speaker B's inability to "*walk*" rather than the fact that they aren't

"permitted to walk inside the building." For this reason, we decided to also resolve the object phrase as a part of our system.

Example 7: Drawback of Transferring English VPE to Hindi VVPE:

7.1) English: A - Can I **walk inside this building**? I'm afraid you *can't*.

7.2) English (With antecedent head verb resolved): A - Can I **walk inside this building**? I'm afraid you *can't walk*.

7.3) Hindi Translation after VPE-VVPE Transfer: - A - kya main **iss imarat ke andar chal sakta hun**? B - mujhe dar hai ki tum **chal** nahi *sakte*.

We also observed that tag questions work very differently in English and in Hindi. Tag questions are usually used to ask for confirmation; however, English makes use of two more types of tag questions that do not play the role of asking for confirmation. These types of tag questions are triggered by modal auxiliaries. The first kind involves an imperative sentence followed by a tag question, and the second kind involves a sentence that uses a volitive modality (e.g., "*want to*", "*wish to*", etc.) and the tag question that follows is the speaker asking for permission to fulfil their desire.

On the contrary, in Hindi, tag questions are only used for asking for confirmation. In Hindi, the tag question is always "*hai na?*", which is equivalent to the tag question "*right?*" in English. The VPE cases triggered by tag questions that ask for confirmation do not need to be resolved, as we observed that their translations aren't erroneous, and resolving them also led to clumsy and disfluent translations. Therefore, we chose not to resolve them. We chose not to resolve tag questions followed by imperative sentences, even though they led to erroneous translations, as resolving them also led to disfluent translations. This is because, in Hindi, an imperative sentence must stand on its own and without any succeeding tag question. We only need to resolve tag questions that follow a sentence with a volitive. Example 8 illustrates this. (Licensors are indexed in the subscript and are demarcated in italics, its corresponding antecedent is in bold and has the same index as the licensor). Here, resolving *won't*₁ gives us a clumsy construction in Hindi, resolving *shall*₂ gives us an incorrect translation and resolving *can*₃ gives us an improved translation.

Example 8: Effect of VPE resolution on translations of tag questions:

8.1) English (Unresolved): A - He will **come to the party**₁, *won't*₁ he <>? B - Yes, he's coming to the party. Now, let us **leave**₂, *shall*₂ we <>? A - Well, I want to **drink there**₃, *can*₃ I <>?

8.2) Hindi Translation (Unresolved): A - vah party mein aaenge, hai na? B - haan, vah party mein aa rahaa hai. ab, chalo chalein, kya ham? A - accha, main vahaan peena chahta hun, kya main kar sakta hun?

8.3) English (Resolved): A - He will come to the party, *won't*₁<**he come to the party**>? B - Yes, he's coming to the party. Now, let us leave, *shall*₂ we <**leave**>? A - Well, I want to drink there, *can*₃ I <**drink there**>?

8.4) Hindi (Resolved): A - vah party mein aaenge, kya vah party mein nahi

aaenge? B - haan, vah party mein aa rahaa hai. ab chalo chalein, chalein? A - accha, main vahaan peena chahta hun, kya main vahaan pee sakta hun?

Antecedent Head Verb Candidate Scoring Post detecting a VPE licensor, we consider all the verbs preceding the licensor in the sentence where VPE was observed as candidates for the head verb antecedent. We do so because in a conversational context, the antecedent is mostly observed in the preceding context of the conversation. Along with that, we also consider all the verbs occurring in the preceding 3 sentences, relative to the sentence with VPE, within the conversation as potential antecedent head verb candidates. Note that if the licensor is part of a tag question, we do not need to look at the previous sentences for antecedent verb candidates. We remove all the auxiliary verbs from the list of candidates. We also remove the plural imperative verb "*let*" from the list. We also eliminate some verb candidates based on the VPE licensor category. In the case of *be* verb licensors, we only keep gerund and past-participle verb candidates with *be* auxiliaries. For *be* verb licensors, we also eliminate all the main verbs belonging to the *be* verb class. For *to* licensor, we remove its parent verb from the list of candidates.

We then score these candidates by comparing their nominal subjects with that of the nominal subject of the VPE licensor. For the candidates present in the same sentence wherein the licensor is detected, we award 2 points to the verb candidate if the nominal subject is exactly the same as that of the licensor's nominal subject. If they do not match, we award 1 point to the candidate if both the nominal subjects are proper nouns and award 1 point if their plurality matches. We penalise the verb candidate by 2 points if there is a mismatch in the passivity of the nominal subjects.

Next, we award scores to verb candidates from sentences preceding the sentence with VPE. Within two-person conversations, we see an abundance of the usage of first-person and second-person pronouns. With this observation from our training data, we awarded scores to the candidates based on the speaker of a sentence when we see first-person or second-person pronouns at the nominal subject positions of the candidate and the licensor. Table 2 from the Appendix section summarizes this scoring system for candidates occurring in previous sentences. The scoring based on the presence of proper nouns, plurality and passivity mentioned above are applied after this step.

For candidates present in the same sentence as the licensor, we penalise a candidate if it is a complement child of the licensor. When assigning scores to candidates from preceding sentences, if the licensor is a *do* verb, then we need to give preference to the candidate that is the oldest ancestor in the list of verbs that are connected by a complement dependency relation. We penalize the oldest ancestor's children by 1 point to do so. For e.g., in "A - *Laura told me today that she has a friend with a car for sale.* B - *Oh, she did <>?*", the licensor is "*did*", and its corresponding antecedent head verb is "*told*", not its complement child "*has*". This rule will penalise "*has*" in the above example.

We award 1 point to candidates whose auxiliary class matches the licensor’s class. For modal auxiliaries, however, we look for exact matching modal auxiliaries to award this point. We also award 1 point to the first verb we obtain by backtracking up the dependency tree from the licensor. Finally, if the licensor is a modal auxiliary verb, we award 1 point to imperative verb candidates. The candidate with the highest score is chosen. In cases where there is a tie in the score, we choose the candidate nearer to the licensor.

Predicate Resolution And Reconstructing The Conversation We would like to address a special case of ellipsis wherein only the non-verbal predicate children are elided. This is observed for *be*, *do* and *have* licensors since they can take the role of a main verb. While studying our training data, we noted that the ellipsis of a non-verbal predicate affects translations only when the licensor is a *be* verb. Hindi’s VVPE property allows *do* and *have* verbs to remain stranded without their non-verbal predicate children in cases where they as auxiliary licensors to elided non-verbal predicates. Therefore, we don’t need to resolve them. The same is not true for *be* verbs with elided non-verbal predicates. These cases were ignored by N. Bafna. et al.’s system [5].

In our training set, we noticed that the antecedent non-verbal predicate of a *be* verb is either present in the same sentence as that of the *be* licensor or in its immediately preceding sentence. Therefore, for resolving such cases, our system restricts itself to these two sentences and looks for verb candidates in them. Here, we only need to resolve the verb candidate’s non-verbal predicate children and not the main verb itself, as the *be* verb licensor doesn’t act as an auxiliary. We only keep the candidates that are either *be* main verbs or are main verbs without any auxiliary licensor. Gerund verbs are also removed as candidates because they act as nouns. Amongst these, we choose the candidate nearest to the licensor in case of a tie. The non-verbal predicates of the chosen candidate are then used for resolution. This resolution is carried out in one of the two following cases: 1) If we see that no main verb antecedent candidates remain after we eliminate verbs from the list of candidates, or 2) If all main verb antecedent candidates have negative scores.

Once the head verb or the parent verb to an elided non-verbal predicate is resolved, we recursively extract its children from the dependency parse tree from the sentence containing the antecedent head verb. The conversation is then reconstructed with the antecedent verb phrase at the site of VPE. If the antecedent head verb has any auxiliary verbs that do not match the licensor’s category, then they are appended while reconstruction. We also change the tense of the antecedent head verb based on the licensor. Note that the tense doesn’t need to be changed for *be* licensors. If the predicate children make use of first-person and second-person pronouns, they are modified after comparing the speaker who utters the sentence with VPE and the speaker who utters the antecedent VP. Based on the presence of negation and what type of sentence the VPE occurs in (question or a normal sentence), we decide the correct position of the antecedent VP and then append it there.

Sample Run Of Resolution Rules Consider Example 6 again for resolution. Our licensors are to_3 , $Will_9$, $will_{10}$, and is_{11} from sentences 6.2, 6.10, 6.11, and 6.13 respectively. For licensor to_3 , the candidate list that we get after eliminating auxiliary verbs and its parent is {"sit"}. Even if "sit" is the only remaining candidate, our system will still award it 2 points based on the scoring system in Table 2 and because of the matching plurality of nominal subjects. VP "sit down" is eventually resolved after to_3 . For both $Will_9$, $will_{10}$, we get a candidate list consisting of {"sit", "listen", "cook"}. Here, in both cases "cook" ends up getting an extra point over the remaining candidates at the end because their auxiliary exactly matches the licensor, and hence is chosen as the antecedent. For the licensor is_{11} , our list of candidates for head verb resolution is empty as we see no gerunds and past participle verb forms with *be* auxiliaries. This triggers the search for a candidate that would act as a parent to an elided non-verbal predicate. Sentence 6.13 doesn't have any other verb, and 6.12 contains the verbs "Cooking", "looks". "Cooking" is removed as it is a gerund. We end up a sole candidate "looks", whose non-verbal predicate child "easy" is resolved after the negation "n't" in Sentence 6.13. The resolved sentences would finally look like this:

Sentence 6.2 (Resolved): I don't want to sit down.
 Sentence 6.10 (Resolved): Will you cook dinner?
 Sentence 6.11 (Resolved): OK, I will cook dinner.
 Sentence 6.13 (Resolved): It most definitely isn't easy.

4 Results and Performance analysis

With our rule-based system, we achieved a precision of 98.835% and a recall of 97.277% for the VPE detection sub-task on our training set. For our test set, we achieved a precision of 93.75% and a recall of 91.263%. We observe that out of 1114 cases of endophoric VPE from our training set, we only need to resolve 888 cases. This is so because from the remaining 226 cases, 39 of them comprise cases triggered by non-auxiliary licensors, and 187 of them are triggered by tag question licensors that need not be resolved. For these 888 cases, our system accurately resolves 75.623% of the data from our training set. Similarly, from our test set, we only need to consider 452 out of 542 VPE cases. For these 452 cases, our system accurately resolves 64.78% of the VPE instances. We now analyse errors that we observed in our rule-based system.

4.1 Error Analysis

We notice a few cases wherein the erroneous resolution is a result of the disagreement between CoreNLP [15] and Spacy [16] dependency parser outputs. In the following example, Spacy [16] incorrectly marks "as" (underlined) as a prepositional child of the licensor "am". Due to this, it is not marked as a licensor, and therefore this error propagates to the resolution sub-task as well.

Example 9: Error propagation due to Dependency parser disagreement:

A - Are you **going shopping today**_{unresolved}? B - I *amundetected* <>, as a matter of fact.

We observe that the comparison of the nominal subject for candidate scoring is a weak criterion to resolve VPE. Consider Example 10 (L is the licenser, C is the correct antecedent head verb, and I is the incorrectly resolved head verb. Their corresponding nominal subjects are marked with *nsubj* in subscript. The text in bold is the correct verb phrase that needs to be resolved). Here, we observe that the subject of the licenser is "We"; however, the correct antecedent head verb "bring"'s nominal subject is "you". Our rule-based system ends up giving a higher score to the incorrect antecedent head verb "watch" because its nominal subject exactly matches the licenser's nominal subject, and it ends up getting 2 points. The candidate "bring" gets 1 point based on the scoring system mentioned in Table 2 of our Appendix. Since "you"'s plurality doesn't match that of the licenser's nominal subject, it doesn't get any more points. This leads to erroneous resolution.

Example 10: Resolution error caused by nominal subject-based scoring system:

A - Well, where I grew up, we_{nsubjI} saw_I movies at a drive-in theatre in our car with the whole family. B - That's cool. I bet you_{nsubjC} could **bring_C your own food**. A - We_{nsubjL} *did_L* <>.

We also observe that in cases where an imperative verb should be resolved, our system ends up giving a higher score to another candidate because the auxiliary class of licenser and the incorrectly resolved candidates match. Consider Example 11 for this. Here, the imperative verb "Study" is the correct antecedent; however, our system resolves the head verb "help" because its auxiliary verb "will" (underlined) is the same as that of the licenser and it ends up receiving an extra point. Also note that here, both candidates will receive 1 point after comparing nominal subjects.

Example 11: Resolution error caused in cases where the antecedent is an imperative verb:

A - **Study_C regularly and diligently**. This_{nsubjI} will help_I you get a good grade. B - I_{nsubjL} *will_L* <>.

For *be* licensers, our rule-based system first tries to resolve the main verb antecedent before searching for the parent verb to the elided non-verbal predicate of a copula/*be* main verb. The system only starts resolving elided non-verbal predicates if no main verb candidates remain or if the remaining verb candidates have negative scores. Because of this, we notice the error that our system ends up incorrectly resolving an antecedent main verb at the site of the ellipsis. In Example 12, our system ends up resolving the gerund verb "walking" at the site of ellipsis instead of the elided non-verbal predicated "that bad" as it has a positive score. (Here, the parent verb to "that bad" is demarcated with an underline.)

Example 12: Resolution error due to higher preference to main verb resolution:

A - You probably don't want to be walking_I around after dark. B - It can't be that bad_C. A - I wish it *weren't*_L <>, but there is actually a lot of crime and prostitution around here.

We also observe incorrect resolutions as we give higher preference to the candidate that is closer to the licenser when we see candidates with a matching score. In Example 13, our system ends up incorrectly resolving the imperative verb "cut" instead of the imperative verb "be" since it is closer to the licenser. Since our licenser is a modal auxiliary verb, both these candidates get a score of 1 for being imperative in nature, but because of closer proximity, "cut" is incorrectly resolved.

Example 13: Resolution error due to the selection of nearest candidate in case of tied scores:

A - **Be**_C careful of the peeler. Don't cut_I your fingers. B - I *will*_L <>.

4.2 Performance On Unseen Data

In our study, we also manually tested the performance of our system on 50 new instances of conversational data from DialyDialog [4] and 50 new instances taken from WSJ corpus that were annotated by Bos and Spenader [2]. All the instances were chosen at random and contain endophoric VPE instances triggered by auxiliary verbs and the infinitival *to*. Within the conversational data, we observed a total of 64 instances of VPE, out of which our system failed to detect 3. For the resolution task, our system successfully resolved 47 out of the 64 instances of VPE, thereby accurately resolving VPE for 73.437% of the VPE instances. This number is comparable to the performance of our system on our annotated data. Within instances taken from the WSJ corpus, we observed a total of 50 instances of VPE, out of which our system failed to detect 11. Our system successfully resolves 56% of the instances correctly. These observations were expected. This is so because our system ignores VPE triggered by Do-So anaphora. It also ignores antecedent candidates that succeed the licenser and also utilizes rules that are dependent on the speaker. We also observe incorrect resolutions due to the penalisation in cases of passivity mismatch of the nominal subject.

5 Conclusion And Future Work

Thus, we present a study on detecting and resolving VPE in the context of handling English-Hindi MT outputs. We present a rule-based system that could be utilized to process English VPE prior to the MT process. We also present novel data on conversational VPE that we used for our study.

The focus of future work can revolve around three primary entry points. First would be testing the scalability and applicability of the rule-based system on text pre-editing, for example, in the case of texts for MT that are transcripts of spoken data, like audiovisual translations or translations of transcripts of live events/broadcasts/lectures. With this, we can also study whether rule-based pre-editing intervention may improve the output in the VPE aspect and to what extent it benefits the process of post-editing translations. Secondly, we observe that the main problem we face comes from the sparsity in gold annotated data. Even the data provided by Bos and Spenader [2] consists of only 487 instances of VPE. We argue that one may approach the problem of VPE resolution as a sequence-to-sequence generation task with training data consisting of several instances of VPE. A rule-based system like ours could be utilized for data augmentation in order to create sufficient data for approaching VPE detection and resolution as a sequence-to-sequence generation task. This would allow us to utilize an LLM (Large Language Model) based approach, where the model 'learns' to resolve problematic sentences prior to translation in the source text. One can then also do a comparative study between a rule-based approach and an LLM-based approach. Finally, we are yet to see how discourse-level English-Hindi MT systems handle instances of VPE. We leave these points to future investigation.

References

1. Shahabi, M., Baptista, J.: A Corpus-Based Translation Study on English-Persian Verb Phrase Ellipsis. *BULAG: Bulletin de Linguistique Appliquée et Générale*, 197–217 (2012)
2. Bos, J., Spenader, J.: An annotated corpus for the analysis of VP ellipsis. *Lang Resources & Evaluation*, vol. 45, pp. 463–494. (2011). <https://doi.org/10.1007/s10579-011-9142-3>
3. Dalrymple, M., Shieber, S.M., Pereira, F.C.N.: Ellipsis and higher-order unification. *Linguist Philos*, vol. 14, pp. 399–452. (1991). <https://doi.org/10.1007/BF00630923>
4. Li, Y., Su, Hui., Shen, X., Li, W., Cao, Z. Niu, S.: DailyDialog: A Manually Labeled Multi-turn Dialogue Dataset. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995. Asian Federation of Natural Language Processing, Taipei, Taiwan (2017)
5. Bafna, N., Sharma, D.: Towards Handling Verb Phrase Ellipsis in English-Hindi Machine Translation. In: *Proceedings of the 16th International Conference on Natural Language Processing*, pp. 150–159. NLP Association of India, International Institute of Information Technology, Hyderabad, India (2019)
6. Hardt, D.: An Algorithm for VP Ellipsis. In: *30th Annual Meeting of the Association for Computational Linguistics*, pp. 9–14. Association for Computational Linguistics, Newark, Delaware, USA (1992)
7. Hardt, D.: An Empirical Approach to VP Ellipsis. *Computational Linguistics*, **23**(4), 525–541. (1997)
8. Nielsen, L. A.: A Corpus-Based Study of Verb Phrase Ellipsis Identification and Resolution. Ph.D. thesis, King's College, London, United Kingdom (2005)
9. McShane, M., Babkin, P.: Detection and Resolution of Verb Phrase Ellipsis. *Linguistic Issues in Language Technology*, **13** (2016)

10. Liu, Z., Pellicer, E. G., Gillick, D.: Exploring the steps of verb phrase ellipsis. In: Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016), pp. 32–40. Association for Computational Linguistics, San Diego, California, USA (2016)
11. Kenyon-Dean, K. Cheung, J.C.K., Precup, D.: Verb Phrase Ellipsis Resolution Using Discriminative and Margin-Infused Algorithms. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1734–1743. Association for Computational Linguistics, Austin, Texas, USA (2016)
12. Aralikkatte, R., Lamm, M., Hardt, D., Søgaard, A.: Ellipsis Resolution as Question Answering: An Evaluation. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 810–817. Association for Computational Linguistics, Online (2021)
13. Manetta, E.: Verb-phrase ellipsis and complex predicates in Hindi-Urdu. *Nat Lang Linguist Theory* vol. 37, pp. 915–953 (2019). <https://doi.org/10.1007/s11049-018-9429-9>
14. Leech, G.: 100 Million Words of English: The British National Corpus (BNC). *Second Language Research* **28**, 1–13. (1992)
15. Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 55–60. Association for Computational Linguistics, Baltimore, Maryland, USA (2014)
16. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017)
17. Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., Zettlemoyer, L.: AllenNLP: A Deep Semantic Natural Language Processing Platform. In: Proceedings of Workshop for NLP Open Source Software (NLP-OSS), Association for Computational Linguistics, Melbourne, Australia (2018)
18. Khullar, P.: Are Ellipses Important for Machine Translation? *Computational Linguistics*, **47**(4), 927–937. MIT Press, Cambridge, Massachusetts, USA (2021). https://doi.org/10.1162/coli_a_00414
19. Voita, E., Sennrich, R., Titov, I.: When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1198–1212. Association for Computational Linguistics, Florence, Italy (2019). <https://doi.org/10.18653/v1/P19-1116>

6 Appendix

In the below tables, **L** refers to the Licensor of VPE, and **A** refers to its respective antecedent.

Table 1. Annotation Schema Summary

Annotation Category	Annotation Label	Notes
ellipsis_where	1 (L and A in the same sentence) 2 (A is in one of the sentences that precedes the sentence containing L) 3 (Sentence with L is a question and its A is in a previous sentence) 4 (L is part of a tag question and its A is present in the same sentence)	This annotation category marks the type of sentence where the licensor is present and where its antecedent is coming from.
antecedent_exists	0 (Exophoric VPE) 1 (Endophoric VPE)	This annotation category marks whether the VPE is Exophoric or Endophoric
ellipsis_type	1 (L is a <i>be</i> verb) 2 (L is a <i>do</i> verb) 3 (L is a <i>have</i> verb) 4 (L is a modal auxiliary) 5 (L is infinitival <i>to</i>) 6 (L is not an Auxilliary)	This annotation category marks the category of the licensor
vpe_sen_index	Sentence index of the sentence where VPE is observed	Sentence Indexing starts from 0
vpe_token_index	Token index of L within the sentence where VPE is observed	Token Indexing starts from 0
ant_sen_index	Sentence index of the sentence from which A is resolved.	Sentence Indexing starts from 0
ant_token_start_index	Starting Token index of A verb phrase	Token Indexing starts from 0
ant_token_end_index	Ending Token index of A verb phrase	Token Indexing starts from 0

Table 2. Nominal subject based scoring system for previous sentence verb candidates

Speaker uttering L vs. Speaker uttering C	L's Nominal Subject Category	C's Nominal Subject Category	Score awarded
Same Speakers	1st-person pronoun	1st-person pronoun	1 2 (On exact match)
Same Speakers	2nd-person pronoun	2nd-person pronoun	1 2 (On exact match)
Different Speakers	1st-person pronoun	1st-person pronoun	0 -1 (when L is a <i>be</i> verb)
Different Speakers	2nd-person pronoun	2nd-person pronoun	0 -1 (when L is a <i>be</i> verb)
Different Speakers	1st-person pronoun	2nd-person pronoun	1
Different Speakers	2nd-person pronoun	1st-person pronoun	1
Speaker is Irrelevant	3rd-person pronoun or Noun	3rd-person pronoun or Noun	1
Speaker is Irrelevant	3rd-person pronoun or Noun	1st-person pronoun	-1
Speaker is Irrelevant	3rd-person pronoun or Noun	2nd-person pronoun	-1
Speaker is Irrelevant	1st-person pronoun	3rd-person pronoun or Noun	-1
Speaker is Irrelevant	2nd-person pronoun	3rd-person pronoun or Noun	-1

Lost in Innu-Aimun Translation - Re-defining Neural Machine Translation for Indigenous Interpreters and Translators Needs

Antoine Cadotte¹, Anne-Christina Thernish^{2,3}, and Fatiha Sadat¹

¹ Université du Québec à Montréal, Montréal (Québec) H3C 3P8 Canada
{cadotte.antoine,sadat.fatiha}@uqam.ca

² Cégep de Sept-Îles, Sept-Îles (Québec) G4R 5B7 Canada
annchristinat@hotmail.com

³ Uashat Mak Mani-Utenam Community

Abstract. Innu-Aimun, one of the most spoken Indigenous languages in Canada, faces significant transmission challenges. Although there is a notable body of Innu-Aimun literature, there is generally not enough documentation written in Innu-Aimun for daily use, just as there are not enough translators and interpreters for the language. We present here collaborative work between Innu-Aimun translators and researchers in computational linguistics to develop translation assistance tools, with the aim of helping language revitalization and preservation. We detail our common position on how should technological assistance tools be developed for Innu-Aimun, which emphasizes the importance of involving Innu translators throughout the entire process and making sure to address language-specific needs. This position is elaborated from joining our respective perspectives (researchers and Innu community member) and expertise (computational linguistics and Innu-Aimun translation). In this spirit, we present preliminary results for the first ongoing steps towards building a first Innu-Aimun - French Neural Machine Translation model. We focus on our participatory process to create aligned parallel corpora and present first results and analyses.

Keywords: Indigenous Language · Innu-Aimun · Corpus Alignment · Neural Machine Translation · Collaboration · Indigenous Translation

1 Introduction

"Innu-Aimun" literally means the language of the Innu [26]. The Innu are a First Nation in Canada, whose members live for the most part in a dozen communities in the north-east of the province of Quebec and in Labrador [2]. One of more than 70 Indigenous languages in Canada, Innu-Aimun was spoken in 2021 by 11,605 locutors (including Naskapi), making it the 4th most spoken in the country. This number has been decreasing in recent years ⁴. And yet, there has been for

⁴ Statistics Canada: Indigenous languages in Canada, 2021. <https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2023029-eng.htm>

several decades a growing body of Innu literature written both in Innu-Aimun and French, a movement that takes its origins from oral stories [34]. Overall, the situation of Innu-Aimun has been described as «*alive but still fragile*» [2].

Although there are a few existing technological tools for Innu-Aimun, such as the online trilingual dictionary [26], there is currently no available tool that takes advantage of more recent or advanced natural language processing methods, or that offers a greater level of assistance or automation. In this context, we are pursuing research work to develop Innu-Aimun translation assistance tools. This work is still at the initiation stage, as we focus on studying feasibility and building basic building blocks such as aligned bilingual corpora.

As noted in the MIT Technology Review series, "AI is enriching a powerful few by dispossessing communities that have been dispossessed before" [12]. Furthermore, Keoni Mahelona mentions that "Data is the last frontier of colonization" [13] and without collaboration with the Indigenous community and putting in the hands of the community, the linguistic tools created for the revitalization of their language, this goal will remain an illusion. The Two-Eyed AI approach has also stated the principle advanced by Mi'kmaw Elders, Albert and Murdena Marshall, to consider ethical issues towards Indigenous knowledge, and their management in tandem with "Western knowledges" or the advancement and deployment of AI systems [5].

There are two main contributions in this paper. First, we present our ongoing collaborative work with Innu translators, to create reference bilingual sentence alignments for Innu-Aimun and French. The goal of this work is to allow the human evaluation of alignment methods for available bilingual Innu-Aimun texts, and help create larger aligned corpora for the development of translation assistance tools with a close collaboration with the community. Second, we offer our position on the development of Innu-Aimun translation assistance tools to help with language revitalization and preservation, and on the involvement of Innu translators in the process, with respect to the way Indigenous research is conducted, by putting the linguistic tool in the hands of the users and community, in each stage of the research and development.

Section 2 presents works related to our research. Section 3 presents our position. Section 4 presents our work on Innu-French alignment. Section 5 explores a few experimental perspectives for the developments to come.

2 Related Works

Presented here are existing technologies for Innu-Aimun, related languages and other Indigenous languages in Canada.

Few technological tools currently exist for Innu-Aimun, and the general state of its language technologies can be considered as nascent [6]. The main technological tool currently available for Innu-Aimun speakers and learners is the trilingual, pan-dialectal online dictionary [26], which allows to search for Innu-Aimun word definitions, translations or usage examples in Innu-Aimun, French

and English. This dictionary is part of a series of web tools proposed for language maintenance [17], which also includes an online verb conjugation application.

There has been projects to create fundamental resources like annotated, aligned and segmented corpora, such as the Innu Language Documentation Project [9] or the Labrador Languages Preservation Database ⁵, but no such resource is currently openly available. More recently, an Innu-Aimun morphological segmenter based on deep-learning was proposed [35]. Fundamental resources like morphological models for segmentation exist for related languages, namely East Cree [1] and Plains Cree [33] [15]. Such resources could potentially be adapted to Innu-Aimun.

More advanced technological tools have been developed for related language. Developments have been presented for Plains Cree word completion [21] and speech synthesis [14], among others. Beyond related languages, developments have been made for several other Indigenous languages in Canada, from very low-resourced to higher resourced languages. A first automatic segmenter was recently proposed for Inuinnaqtun, one of the most endangered Indigenous languages in Canada, [23]. The most resourced among Indigenous languages in Canada is Inuktitut, whose official status in the Nunavut legislative assembly has allowed the development of an aligned corpus counting over 1 million sentence pairs, and a Neural Machine Translation (NMT) model reaching high translation scores [16]. Microsoft has since included Inuktitut in its translator⁶. This high availability of bilingual data has also allowed further research like the improvement of NMT and segmentation results [30] [24] and the study of gender bias in Inuktitut NMT models [11] [22].

3 Our Position: Developing Collaboration and Translation Tools to Help Revitalize Innu-Aimun

We present here our position on a multidisciplinary and community-driven project dedicated to the development of translation tools to help revitalize Innu-Aimun. This position reflects our observations and fieldwork, along with insights from working closely with Innu-Aimun, with the Indigenous community of Ushat Mak Mani-utenam, as well as experience in our respective fields (computational linguistics and Innu-Aimun translation).

3.1 The Role of Innu-Aimun Translation for the Language’s Revitalization

As of today, there are not enough documents available in Innu-Aimun within Innu communities. The vast majority of them are in French only. This not only

⁵ Labrador Languages Preservation Database. <https://www.hss.mun.ca/research/showcase/labrador-languages-preservation-database/>

⁶ Microsoft News - Microsoft Introduces Inuktitut to Microsoft Translator (January 27, 2021). <https://news.microsoft.com/en-ca/2021/01/27/microsoft-introduces-inuktitut-to-microsoft-translator/>

includes the documents and websites of provincial and federal governments, but also local community governments. We believe that this does not encourage community members to learn or use the language. We believe that if there were more documents available in Innu-Aimun for the population, it would provide more opportunities to use the language, to learn it or to become more fluent. In that sense, the work of Innu-Aimun translators could be very important to help improve language learning and usage within the community.

Although there has been increasing demand for Innu-Aimun translation, the number of professional translators/interpreters⁷ is insufficient. This shortage has been documented by the Viens Commission, not only for Innu-Aimun but for other Indigenous languages in the province of Quebec⁸. Recently, a new program for Innu-Aimun translation and interpretation has been created at Cégep de Sept-Îles⁹. Students set to graduate this year form the first cohort. Members of this cohort that have participated in the present project have observed that they are increasingly seen within the community as references and ambassadors for Innu-Aimun, and as a source of inspiration, just like translators are in general.

We believe tackling the lack of freely available text translated in Innu-Aimun is a top priority for Innu-Aimun revitalization. In view of the demand for translators and of the role they play for the community, translation assistance tools should be considered among the most relevant technological tools to develop for Innu-Aimun. Furthermore, we believe that technological tools for Innu-Aimun translation could be useful not only to professional translators, but to the broader community. Assisting people in translating to Innu-Aimun could motivate them to use it more often. Having an automated translation assistance tool available as a smartphone app could foster the Innu youth's interest in trying to improve their fluency in the language or to learn it.

3.2 How to Develop Tools to Assist Innu-Aimun Translators?

From the beginning to the end of the development process, tools should address Innu-Aimun translators' needs as expressed by them. This principle is applicable to any translation tool, but in the case of Innu-Aimun it can serve to address characteristics of the language that differentiate it from European languages or other higher-resourced languages.

For example, the polysynthetic nature of Innu-Aimun and related translation difficulties can be taken into consideration. Since Innu-Aimun words are often formed by joining morphemes together, the rules to properly do so and to form words equivalent to French ones can be hard to master and apply. As such, tools

⁷ N.B. Translation and interpretation in the community is most often performed by the same individuals.

⁸ Final report of the Viens Commission (in French). https://www.cerp.gouv.qc.ca/fileadmin/Fichiers_clients/Rapport/Rapport_final.pdf

⁹ Cégep de Sept-Îles launches a new program for Innu-Aimun translation and interpretation (October 18, 2021 - in French). <https://www.lemanic.ca/2021/10/18/le-cegep-de-sept-iles-lance-un-programme-de-traduction-en-langue-innue/>

integrating this morphological aspect (e.g. integrating a morphological analyser within an assistance tool) would be more useful to Innu-Aimun translators.

Other examples of language-specific aspect we believe should be incorporated in the development of assistance tools are: the revitalization of older or less used words¹⁰, the inclusion of dialectal variations, as well as the inclusion of traditional or ancestral knowledge.

Assistance tools could take several different forms, varying in development complexity and data amount requirements. With the current amount of bilingual text available [6], one of the lowest hanging fruits could be to extend the search function of the online dictionary to multi-word French expressions. Resulting Innu-Aimun words could be ordered by relevance for the translation of the entire expression or sentence. Other, more difficult developments include translation memory and machine translation. Instead of insisting on the usefulness of one tool over the others, we believe they should be seen as complementary. Their development should be ordered from short to long term goals, reflecting the amount of data and research work needed. In addition, priority should be given to the development of common building blocks, such as the collection and annotation bilingual texts or the development of morphological segmentation.

Whatever the tool or the development stage, continuous feedback and evaluation should be undertaken in collaboration with Innu-Aimun translators. Assistance tool evaluation should not only take the form of evaluation metrics on test sets, it should also include qualitative feedback. Translators can further participate by providing post-editings, or by testing the tool interactively, using a human-in-the-loop method (see section 5).

4 Creating Aligned Corpora in Collaboration with the Innu-Aimun Translators

We present here collaborative work to create aligned bilingual corpora. These corpora are created from texts that were published both in Innu-Aimun and French. The sentence pairs from those texts are aligned in collaboration with students in Innu-Aimun translation.

4.1 Community-Driven Approach

Our collaborative and community-driven approach is derived from the position specified in the previous section, but is also based on the imperative of community involvement which has been noted more broadly within the computational linguistics research community when it comes to lower resourced and indigenous languages. As an example, Bird [3] draws upon his experience in multiple projects, as well as other studies, to highlight the negative impact that a colonial approach to language technology development can have on an indigenous

¹⁰ There has been lexical erosion in some communities due to high rates of bilingualism [7]

community, emphasizes the importance of the involvement of the community and gives keys to a decolonial approach [3]. Successful examples of participatory research in NLP exist, such as the Masakhane project for African languages [29]. Some computational linguistics tools for learning have also been evaluated in part with native speakers, such as *néhiyawétân* [4] for which feedback was collected from usage of the tool and noted its usefulness for users.

Practically speaking, and for the first steps of our project, our approach has been defined based on the following principles. First, students in Innu-Aimun translation should have personal benefits in participating in the research. Second, the Innu students' time in the translation program is precious and it should not be taken from them, as it is limited.

Consequently, the corpus alignment activity takes the form of an extracurricular translation practise offered to students, independently of their program. Participation is on a voluntary basis, with financial compensation.

Alignment is done using a simple online shared document, in which participants can write alignments manually. Figure 1 presents an example of the form used by participants. Bilingual texts have first been extracted from their PDF version and cleaned, paragraphs separated and sentence numbered. Since automated extraction, cleaning and paragraph alignment sometimes generates errors, the work done by participants can also serve to highlight those mistakes.

Français	Innu-Aimun	Alignement	Note
[0] À mes huit enfants. [1] Préambule	[0] Nitouassimot umenu nishuauush etseshiht. [1] Tshitshipanu aimun	0.0 1.1	
[0] Je remercie chacun de ceux qui m'ont aidée à faire ce livre que j'ai fait. Et je serais heureuse de voir d'autres Indiens écrire, en langue indienne.	[0] Ume mashinaikan ka tutaman ka itashihit ka uitsihit tshetshi tutaman ninashikumauat kassinu. [1] Kie nipa minuemitentshetshi uapetaman kutak innu tshetshi mashinaishet e innushient.	0.0 0.1	

Fig. 1. Sample from the alignment document used by participants

Using a simple document where one can write free text—instead of using a professional tool—allows participants to freely express what they think the alignment should be, regardless of whether a standard alignment tool or method would permit it. This allows to identify limits of existing alignment methods in the context of bilingual Innu-Aimun texts.

The document also provides space for personal notes, where participants can express opinion on the text or paragraph, identify inexact or non-standard translations. They can also make note of alignments that are more complex and express those alignments in a more qualitative manner.

In addition to the alignment document, participants are invited to openly share their thoughts on the text they are asked to align, either orally or through comments directly written in the document's 'Notes' sections. Such thoughts can be helpful to learn about qualitative aspects of the text that can be evasive to non-speakers of Innu-Aimun. For example: is the text written in a standard

language, or in a local dialect? Does the translation generally seem correct? Does the text use a generic vocabulary, or does it use a more specialized or particular one? Is the text biased? Those characteristics can subsequently inform the type of use for the resulting reference alignments (i.e. as training data, as a gold-standard for evaluation etc.).

4.2 Aligned Corpora Analysis

Two bilingual texts have been aligned so far. Both are from Innu author An Antane Kapesh and have been published in bilingual, Innu-Aimun and French editions. Table 1 presents preliminary analysis results for those two corpora, *Eukuan nin matshi-manitu innu-ishkueu* [18] (abbreviated to kapesh-1 in the table) and *Tanite nene etutamin nitassi?* [19] (abbreviated to kapesh-2 in the table).

Table 1. Preliminary analysis of reference corpora aligned in collaboration with Innu-Aimun translators

	Kapesh-1	Kapesh-2	Combined
nb paragraphs	163	149	312
nb sentences	796	484	1280
% non-standard alignments	14.7%	11.4%	13.1%
mean nb of tokens per sentence (French-to-Innu-Aimun ratio)	1.46	1.41	1.44
mean nb of characaters per token (Innu-Aimun-to-French ratio)	1.63	1.58	1.62
vocab size (Innu-Aimun)	2846	1839	4229
vocab size (French)	2472	1856	3514

Table 1 presents the percentage of alignments we consider as non-standard. For example, the translator sometimes considered it could not align sentences properly without dividing them into sub-sentences. Most state-of-the-art alignment methods (e.g. Vecalign [36]) or classic ones (e.g. Moore [28] and Gale & Church [10]) do not support such alignment operation. Table 2 gives an example of a non-standard case, where a sentence split is deemed necessary by the translator. In this example, where italic texts coincide in both languages, the correct way to align would require joining the second Innu-Aimun sentence with the second half of the first. With a method allowing many-to-many alignments like Vecalign, the best the algorithm can do is joining both sentences as one whole, for both languages, but many other methods cannot perform many-to-many operation. We see such cases as opportunities to examine the limits of existing alignment methods.

This preliminary analysis also highlights the polysynthetic nature of Innu-Aimun, with a greater number of token per sentence for French than for Innu-Aimun. Not only do Innu-Aimun words tend to be longer, they can also be

Table 2. Example of "non-standard" alignment

Innu-Aimun	French
#1 Ne Kauitenitakusht katshi minikut nenua auassa tat ^u shuniau- aueshisha, ekue mishta-papit ekue itenitak : « <i>Apu nita tshika ut tshiveian ute katshi takushinian.</i>	Après avoir reçu de l'enfant les four- rures, le Polichinelle éclate de rire.
#2 <i>Ekute ute tshe ut uenutishian ! »</i>	« <i>Maintenant que je suis venu ici, jamais je ne m'en irai, c'est ici que je vais faire fortune ! » se dit-il.</i>

inflected in a great number of ways [8]. This is reflected in the fact that there is a greater number of individual words in Innu-Aimun when combining both corpora than for French. Those findings emphasize the importance of further developing morphological segmentation for Innu-Aimun in order to have useful cross-lingual tools.

5 Experimental Perspectives

One experimental option for the involvement of translators is post-editing, which has been showed to help improve translation in the context of low-resources language pairs. A recent example showed improvements for English to Manipuri and Mizo, a language pair for which there are only a few thousands sentences available [32].

Beyond traditional post-editing, another way of including translators in the process could be using interactive, human-in-the-loop methods or an advanced machine learning approach such as Reinforcement Learning from Human-Feedback (RLHF). For example, one study showed that using this approach to incorporate user translator feedback on partial translations produced by Neural Machine Translation helped improved translation scores and helped reduce the amount of post-editing required by translators [20]. This type of approach could be promising, considering the very low availability of Innu-Aimun translators and interpreters.

Other strategies include the control language or pre-editing, that should be considered in the development of an efficient NMT for Innu-Aimun translators and interpreters, to help eliminate disturbing or misspelled elements and thus afford better quality translations [27]. An example is the spell checker for endangered Languages in Amazonia [31]. Lastly, prompt-driven NMT could be an interesting way to take into consideration the important style variations in the available bilingual texts, as well as dialectal differences. Recently, a *Prompt-Transformer* architecture was proposed in order to incorporate prompts as constraints, which helped increase adequacy and fluency of translations [25].

6 Conclusion

We presented here our position on the crucial involvement of Innu-Aimun translators in the development of translation assistance tools, at every step of the process. This is required not only to evaluate the tools' performance, but also to identify priorities in their development, based on language needs and specificity.

We presented our ongoing collaborative and community-driven work to create reference aligned Innu-Aimun-French corpora and presented a preliminary analysis. This will allow us to subsequently obtain first results for a translation model, determine development feasibility with the current amount of bilingual data available and provide further insights on the future experimental directions to take.

Acknowledgements We would like to thank the following Indigenous contributors to the Innu-Aimun translation and interpretation program at Cégep of Sept-Îles, Québec, for their participation in this research: Roseline Malleck, Sabine Mestkosho and Judith McKenzie. We would also like to acknowledge Monique Durand's contribution to the creation of the Innu-Aimun translation program at Cégep of Sept-Îles and to thank her and her colleagues at Cégep of Sept-Îles in the continuing education department for having helped us in setting this collaboration. Finally, we would like to thank the reviewers for their useful and constructive suggestions.

References

1. Arppe, A., Junker, M.O., Torkornoo, D.: Converting a comprehensive lexical database into a computational model: The case of East Cree verb inflection. In: Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages. pp. 52–56. Association for Computational Linguistics, Honolulu (Mar 2017). <https://doi.org/10.18653/v1/W17-0108>, <https://aclanthology.org/W17-0108>
2. Baraby, A.M., Junker, M.O., Mollen, Y.: A 45-year old language documentation program first aimed at speakers: the case of the Innu (Mar 2017), <http://hdl.handle.net/10125/41973>
3. Bird, S.: Decolonising speech and language technology. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 3504–3519. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). <https://doi.org/10.18653/v1/2020.coling-main.313>, <https://aclanthology.org/2020.coling-main.313>
4. Bontogon, M.A.: Evaluating nêhiyawêtân: A computer assisted language learning (CALL) application for Plains Cree. Ph.D. thesis, University of Alberta (2016), <https://doi.org/10.7939/R3VD6P81C>
5. Bourgeois-Doyle, D.: Two-eyed ai: a reflection on artificial intelligence. The Canadian Commission for UNESCO. United Nations Educational, Scientific and Cultural Organization (2019)

6. Cadotte, A., Le Ngoc, T., Boivin, M., Sadat, F.: Challenges and Perspectives for Innu-Aimun within Indigenous Language Technologies. In: Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages. pp. 99–108. Association for Computational Linguistics, Dublin, Ireland (May 2022). <https://doi.org/10.18653/v1/2022.computel-1.13>, <https://aclanthology.org/2022.computel-1.13>
7. Drapeau, L.: Bilinguisme et érosion lexicale dans une communauté montagnaise. In: Martel, P., Maurais, J. (eds.) *Langues et sociétés en contact: Mélanges offerts à Jean-Claude Corbeil*, pp. 363–376. Max Niemeyer Verlag (2014). <https://doi.org/doi:10.1515/9783110932935.363>, <https://doi.org/10.1515/9783110932935.363>
8. Drapeau, L.: *Grammaire de la langue innue*. PUQ (2014)
9. Drapeau, L., Lambert-Brétière, R.: The Innu Language Documentation Project. In: Proceedings of the 17th Foundation for Endangered Languages Conference (Oct 2013), <https://ir.library.carleton.ca/pub/13609>
10. Gale, W.A., Church, K.W.: A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics* **19**(1), 75–102 (1993), <https://aclanthology.org/J93-1004>, place: Cambridge, MA Publisher: MIT Press
11. Hansal, O., Le, N.T., Sadat, F.: Indigenous language revitalization and the dilemma of gender bias. In: Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP). pp. 244–254. Association for Computational Linguistics, Seattle, Washington (Jul 2022). <https://doi.org/10.18653/v1/2022.gebnlp-1.25>, <https://aclanthology.org/2022.gebnlp-1.25>
12. Hao, K.: Artificial intelligence is creating a new colonial world order. *MIT technology Review* (2022)
13. Hao, K.: A new vision of artificial intelligence for the people. *MIT technology Review* (2022)
14. Harrigan, A., Arppe, A., Mills, T.: A preliminary Plains Cree speech synthesizer. In: Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers). pp. 64–73. Association for Computational Linguistics, Honolulu (Feb 2019), <https://aclanthology.org/W19-6009>
15. Harrigan, A.G., Schmirler, K., Arppe, A., Antonsen, L., Trosterud, T., Wolvengrey, A.: Learning from the computational modelling of Plains Cree verbs. *Morphology* **27**(4), 565–598 (Nov 2017). <https://doi.org/10.1007/s11525-017-9315-x>, <https://doi.org/10.1007/s11525-017-9315-x>
16. Joanis, E., Knowles, R., Kuhn, R., Larkin, S., Littell, P., Lo, C.k., Stewart, D., Micher, J.: The Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 with Preliminary Machine Translation Results. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 2562–2572. European Language Resources Association, Marseille, France (May 2020), <https://aclanthology.org/2020.lrec-1.312>
17. Junker, M.O., Mollen, Y., St-Onge, H., Torkornoo, D.: Integrated web tools for Innu language maintenance. In: Papers of the 44th Algonquian Conference. pp. 192–210 (2016)
18. Kapesh, A.: *Je suis une maudite Sauvagesse Eukuan nin matshi-manitu inushkueu: EUKUAN NIN MATSHI-MANITU INNUSHKUEU*. Mémoire d’encrier (2019), <https://books.google.ca/books?id=JXRmEAAAQBAJ>
19. Kapesh, A.: *Qu’as-tu fait de mon pays? Tanite nene etutamin nitassi?: TANITE NENE ETUTAMIN NITASSI*. Mémoire d’encrier (2020), <https://books.google.ca/books?id= SXFmEAAAQBAJ>

20. Lam, T.K., Schamoni, S., Riezler, S.: Interactive-predictive neural machine translation through reinforcement and imitation. In: Proceedings of Machine Translation Summit XVII: Research Track. pp. 96–106. European Association for Machine Translation, Dublin, Ireland (Aug 2019), <https://aclanthology.org/W19-6610>
21. Lane, W., Harrigan, A., Arppe, A.: Interactive word completion for Plains Cree. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3284–3294. Association for Computational Linguistics, Dublin, Ireland (May 2022). <https://doi.org/10.18653/v1/2022.acl-long.232>, <https://aclanthology.org/2022.acl-long.232>
22. Le, N.T., Hansal, O., Sadat, F.: Challenges and issue of gender bias in under-represented languages: An empirical study on Inuktitut-English NMT. In: Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages. pp. 89–97. Association for Computational Linguistics, Remote (Mar 2023), <https://aclanthology.org/2023.compute1-1.13>
23. Le, N.T., Sadat, F.: Towards a First Automatic Unsupervised Morphological Segmentation for Inuinnaqtun. In: Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas. pp. 159–162. Association for Computational Linguistics, Online (Jun 2021). <https://doi.org/10.18653/v1/2021.americasnlp-1.17>, <https://aclanthology.org/2021.americasnlp-1.17>
24. Le, N.T., Sadat, F.: Towards a low-resource neural machine translation for indigenous languages in Canada. *Traitement Automatique des Langues* **62**(3), 39–63 (2021), <https://aclanthology.org/2021.tal-3.3>
25. Li, Y., Yin, Y., Li, J., Zhang, Y.: Prompt-driven neural machine translation. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 2579–2590. Association for Computational Linguistics, Dublin, Ireland (May 2022). <https://doi.org/10.18653/v1/2022.findings-acl.203>, <https://aclanthology.org/2022.findings-acl.203>
26. Mailhot, J., MacKenzie, M., Junker, M.O.: Online innu dictionary (2013), <https://dictionary.innu-aimun.ca/>
27. Miyata, R., Fujita, A.: Understanding pre-editing for black-box neural machine translation. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics pp. 1539–1550 (2021)
28. Moore, R.C.: Fast and accurate sentence alignment of bilingual corpora. In: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: Technical Papers. pp. 135–144. Springer, Tiburon, USA (Oct 2002), https://link.springer.com/chapter/10.1007/3-540-45820-4_14
29. Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohunge, T., Akinola, S.O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R.A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M.M., Adeyemi, M., Mokgesi-Seling, M., Okegbemi, L., Martinus, L., Tajudeen, K., Degjila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J.T., Abbott, J., Orife, I., Ezeani, I., Dangana, I.A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Espoir, M., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C.C., Dossou, B.F.P., Sibanda, B., Basse, B., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., Bashir, A.: Participatory research for low-resourced machine translation: A case study in African languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 2144–2160. Association for Computational Linguistics, Online (Nov 2020). <https://doi.org/10.18653/v1/2020.findings-emnlp.195>, <https://aclanthology.org/2020.findings-emnlp.195>

30. Ngoc Le, T., Sadat, F.: Revitalization of Indigenous Languages through Pre-processing and Neural Machine Translation: The case of Inuktitut. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 4661–4666. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). <https://doi.org/10.18653/v1/2020.coling-main.410>, <https://aclanthology.org/2020.coling-main.410>
31. Oncevay, A., Cardoso, G., Alva, C., Ávila, C.L., Vásquez Balarezo, J., Rodríguez, S.E., Camaiteri, D.S., Rojas, E.Z., Francis, D.L., Bautista, J.L., Rios, N.A., Cesareo, R.Z., Montoya, H.E.G., Zariquiey, R.: Schaman: Spell-checking resources and benchmark for endangered languages from amazonia. Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 411—417 (2022)
32. Sanayai Meetei, L., Singh, T.D., Bandyopadhyay, S., Vela, M., van Genabith, J.: English to Manipuri and Mizo post-editing effort and its impact on low resource machine translation. In: Proceedings of the 17th International Conference on Natural Language Processing (ICON). pp. 50–59. NLP Association of India (NLP AI), Indian Institute of Technology Patna, Patna, India (Dec 2020), <https://aclanthology.org/2020.icon-main.7>
33. Snoek, C., Thunder, D., Lõo, K., Arppe, A., Lachler, J., Moshagen, S., Trosterud, T.: Modeling the Noun Morphology of Plains Cree. In: Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages. pp. 34–42. Baltimore, Maryland, USA (Jun 2014), <https://aclanthology.org/W14-2205.pdf>
34. St-Gelais, M.: Une histoire de la littérature innue. PUQ, Tshakapesh Institute, Montréal, Canada (2022), <https://www.puq.ca/catalogue/livres/une-histoire-litterature-innue-4253.html>
35. Tan Le, N., Cadotte, A., Boivin, M., Sadat, F., Terraza, J.: Deep learning-based morphological segmentation for indigenous languages: A study case on innu-aimun. In: Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing. pp. 146–151. Association for Computational Linguistics, Hybrid (Jul 2022). <https://doi.org/10.18653/v1/2022.deeplo-1.16>, <https://aclanthology.org/2022.deeplo-1.16>
36. Thompson, B., Koehn, P.: Vecalign: Improved Sentence Alignment in Linear Time and Space. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 1342–1348. Association for Computational Linguistics, Hong Kong, China (Nov 2019). <https://doi.org/10.18653/v1/D19-1136>, <https://aclanthology.org/D19-1136>

Author Index

- Alsolami, Abdelalah, 12
Aquilina, Keith, 203
- Barkarson, Bjarni, 203
Baselli, Valentina, 157
Borg, Claudia, 203
Bouillon, Pierrette, 195, 294
Braun, Sabine, 147, 167
- Cadotte, Antoine, 342
Chang, Su, 114
Chen, Xiaoyu, 282
- Davitti, Elena, 64, 147, 167
Deng, Yadong, 282
Deshpande, Aniruddha, 325
do Carmo, Félix, 26
- Farrell, Michael, 52, 108
Fernandez-Parra, Maria, 12
Fiorini, Susanna, 41
Fishel, Mark, 203
- Geng, Aiju, 282
- Haddad Haddad, Amal, 133
Han, Yu, 282
- Jiang, Yanfei, 114, 282
- Kadar, Fanni, 268
Korybski, Tomasz, 64, 147
- Lalitha Devi, Sobha, 315
Li, Wei, 282
Li, Xiaochun, 282
Li, Yinglu, 114
Liu, Limin, 282
Liu, Xiaoqin, 282
Liu, Yilun, 114
Liyanapathirana, Jeevanthi, 195
López-Arroyo, Belén, 97
- Ma, Wenbing, 114
Macken, Lieve, 41
Macklovitch, Elliott, 124
- Maisto, Alessandro, 209
Meeus, Laurens, 41
Miaomiao, Ma, 282
Migdisi, Kristin, 41
Monti, Johanna, 1
Motika, Željka, 203
Murgolo, Elena, 288
Mutal, Jonathan, 195
- Ning, Xie, 282
- Oliver, Javier, 209
Oncins, Estella, 177
Orăsan, Constantin, 26
- Palenzuela-Badiola, Leire, 227
Peng, Song, 114
- Qiao, Xiaosong, 114
- R K Rao, Pattabhi, 315
Radic, Zeljko, 167, 188
Recski, Gabor, 268
Rescigno, Argentina, 1
Rodríguez González, Eloy, 147
- Sadat, Fatiha, 342
Saeed, Muhammad Ahmed, 147
Sánchez-Gijón, Pilar, 227
Sanz-Valdivieso, Lucía, 97
Serrat-Roozen, Iris, 177
Sharma, Dipti, 325
Spiteri, Donatienne, 203
Sundar Ram, Vijay, 315
Szoc, Sara, 41
- Tadić, Marko, 203
tao, shimin, 114, 282
Taylor-Stilgoe, Eleanor, 26
Tezcan, Arda, 41
Thernish, Anne-Christina, 342
- Ünlü, Cihan, 78
- Vanallemeersch, Tom, 41
Vasiļevskis, Artūrs, 203

Vidrequin, Magali, 305

Volkart, Lise, 294

Wang, Minghan, 114

Yang, Hao, 114, 282

Yang, Jun, 12

Yanqing, Zhao, 114, 282

Zhang, Min, 114, 282

Zhang, Zhaodi, 282

Zhao, Xiaofeng, 114

Zhu, Junhao, 114

Ziediņš, Jānis, 203