Acknowledgements

AsLing wishes to thank and acknowledge the support of the sponsors of TC43:

Gold Sponsors

[Gold Sponsors Logos]

Silver Sponsor

[RWS Logo]
Preface

For over 40 years the Translating and the Computer conferences, organised since 2014 by the International Association for the Advancement in Language Technology (AsLing), have served as a unique forum for academics, developers, users, and vendors of computer aids for translators, of other translation technology tools, and increasingly, for interpreters and others performing new roles in our industry.

In 2021, for the second year, the Covid-19 pandemic prevented people from traveling to London as they had done for all previous sessions until 2019, to meet in person and exchange face-to-face their experience, concerns, challenges and achievements. After weighing the pros and cons of a second virtual conference and waiting until the very last moment to make a final decision, the AsLing Executive Committee requested the Conference Coordinators to organize a fully online event. The result was a success and allowed a very large audience from all around the world to participate and join the circle of traditional in-person delegates for the 21 presentations, 10 workshops and 3 panel discussions covering a broad range of subjects and tools. The 3 chill-out sessions also brought together freelance and in-house translators, interpreters, researchers and businesspeople from translation companies, international organisations, universities and research centres, and offered them opportunities to exchange ideas, and to learn about and discuss the latest developments in translation technologies.

TC43 Online featured speakers and workshop moderators from academia, industry and the professions that provided insights on the latest developments in the language industry. Recordings of all sessions of the conference, augmented in most cases by PowerPoint slides are available on the asling.org/tc43/ website. These Proceedings contain elaborated texts of many presentations as well as extended reports on one of the panels and one of the workshops held during TC43 Online. The Conference chairs are delighted to thank this edition’s prominent keynote and invited speakers: Sharon O’Brien is a professor of Translation Studies in the School of Applied Language and Intercultural Studies, Dublin City University, Ireland. She talked about how AI has become more powerful and integrated into our lives and how it affects translation, trying to answer these questions from an agent, product, and process perspective and proposing a number of actions all stakeholders could consider as we progress further into the human-machine era of translating. Bruno Pouliquen heads the Advanced Technology Application Center at the World Intellectual Property Organization (WIPO) in Geneva, where he has been working since 2009, and is in charge of exploring and applying machine learning techniques to intellectual property applications. He talked about the importance of having access to multilingual information in an international organization and about the use of available libraries and multilingual data. To help users access multilingual information, this organisation has developed its own machine learning tools, such as WIPO Translate, which allows users to read information in 22 different language pairs, and WIPO Speech-to-Text©, which provides textual (and searchable) access to conferences in the six official UN languages. Jean Nitzke is an associate professor for Translation with a focus on Translation Technology at the University of Agder, Norway. She talked about the ways in which the professional translation market was changing due to the ongoing digitalisation and the increasing application of artificial intelligence and presented a revised model for post-editing competences developed to better suit practical aspects. She also presented three potential job profiles that were in line with the model.

We thank all who submitted proposals to the conference and those authors who produced full versions of their papers for these Proceedings, as well as all whose slides and recordings are available on the AsLing website. A special thank-you goes to all the delegates wherever they were, who, by taking part and interacting remotely with presenters and fellow participants, enriched this conference giving living acknowledgement to this special event. We are grateful to the members of the Programme Committee who carefully reviewed the submissions as well as all additional reviewers who helped assess some of the final papers and to our fellow members of the Organising Committee, who played key roles in ensuring that the 43rd conference online took place and linked people from all continents. Last but not least, we thank our sponsors and all those who lent their support, helping to make both the conference and these Proceedings possible.

Conference Chairs
João Esteves-Ferreira, Ruslan Mitkov, Maria Recort Ruiz, Olaf-Michael Stefanov
The Executive Committee of AsLing establishes several bodies each year, to organise and carry out the annual conference. Membership in these bodies overlap. The tables below show membership in these bodies for TC43.

**Conference Organising Committee:**
Denis Dechandon, European Union (Session Chair)
João Esteves-Ferreira, Tradulex (Conference Chair)
Ruslan Mitkov, University of Wolverhampton (Conference Chair)
Maria Recort Ruiz, International Labour Office (Conference Chair)
Vilelmini Sosoni, Ionian University (Session Chair)
Olaf-Michael Stefanov, United Nations (ret.), (Conference Chair)

**Coordinators:** Maria Recort Ruiz and Olaf-Michael Stefanov

**Editors of the Proceedings:**
The Conference Chairs:
João Esteves-Ferreira
Ruslan Mitkov
Maria Recort Ruiz
Olaf-Michael Stefanov

With the assistance of:
David Chambers
Juliet Margaret Macan
Vilelmini Sosoni

**Programme Committee:**
Juan José Arevalillo, Hermes Traducciones
Sheila Castilho, Dublin City University
David Chambers, AsLing Honorary Member
Caroline Champsaur, Organisation for Economic Co-operation and Development
Eleanor Cornelius, University of Johannesburg
Gloria Corpas Pastor, University of Málaga
Joanna Drugan, University of East Anglia
David Filip, CNGL / ADAPT
Camelia Ignat, Joint Research Centre of the European Commission
Raisa McNab, UK Association of Translation Companies
Vilelmini Sosoni, Ionian University
Paola Valli, Project Manager, Tamedia
Nelson Verástegui, International Telecommunications Union (ret.)
David Verhofstadt, European Investment Bank
# Contents

## Section A: MT, MT evaluation and post-editing

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>An Empirical Study of Whether Limiting Lexical Ambiguity Is the Most Effective Way to Improve Machine Translation Accuracy</td>
<td>7</td>
</tr>
<tr>
<td>Hong Xu</td>
<td></td>
</tr>
<tr>
<td>Who needs an MQM Scorecard?</td>
<td>17</td>
</tr>
<tr>
<td>Catherine Marshall Martins and Alan K Melby</td>
<td></td>
</tr>
<tr>
<td>MT Quality and its effects on post-editors and end-users</td>
<td>26</td>
</tr>
<tr>
<td>Maria Stasimioti and Vilelmini Sosoni</td>
<td></td>
</tr>
<tr>
<td>A Report on the TC43 Workshop: Drafting effective machine translation post-editing guidelines</td>
<td>43</td>
</tr>
<tr>
<td>Viveta Gene and Lucía Guerrero</td>
<td></td>
</tr>
<tr>
<td>Integrating post-editing with Dragon speech recognizer: a use case in an international organization</td>
<td>55</td>
</tr>
<tr>
<td>Jeevanthi Liyanapathirana and Pierrette Bouillon</td>
<td></td>
</tr>
</tbody>
</table>

## Section B: Interpreting Technology

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>KUDO Interpreter Assist: Automated Real-time Support for Remote Interpretation</td>
<td>68</td>
</tr>
<tr>
<td>Claudio Fantinuoli, Giulia Marchesini, David Landan and Lukas Horak</td>
<td></td>
</tr>
<tr>
<td>A Paper on the Conference Panel &quot;In-booth CAI Tool Support in Conference Interpreter Training and Education&quot;</td>
<td>78</td>
</tr>
<tr>
<td>Susana Rodriguez, Francesca Maria Frittella and Alicja M. Okoniewska</td>
<td></td>
</tr>
<tr>
<td>ASR-CAI tool-supported SI of numbers: Sit back, relax and enjoy interpreting?</td>
<td>88</td>
</tr>
<tr>
<td>Francesca Maria Frittella</td>
<td></td>
</tr>
<tr>
<td>Development of technological competences: remote simultaneous interpreting explored</td>
<td>103</td>
</tr>
<tr>
<td>Raquel Lázaro Gutiérrez and Gabriel Cabrera Méndez</td>
<td></td>
</tr>
</tbody>
</table>

## SECTION C: Corpora

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepL vs Google Translate: which is the best at translating MWEs from French into Polish?</td>
<td>110</td>
</tr>
<tr>
<td>A multidisciplinary approach to corpora creation and quality translation of MWEs</td>
<td></td>
</tr>
<tr>
<td>Emmanuelle Esperança-Rodier and Damian Frankowski</td>
<td></td>
</tr>
<tr>
<td>Introducing the PETIMOD Corpus: A Resource for the Analysis of Personification in EU Mediated and Non-Mediated Discourse</td>
<td>128</td>
</tr>
<tr>
<td>Fernando Sánchez Rodas</td>
<td></td>
</tr>
<tr>
<td>OCCAM: cross-lingual unlocking of non-digital texts</td>
<td>133</td>
</tr>
<tr>
<td>Laurens Meeus, Joachim Van den Bogaert, Arne Defauw, Oan Stultjens, Sara Szoc, Tom Vanallemeersch, Frederic Everaert and Koen Van Winckel</td>
<td></td>
</tr>
</tbody>
</table>
An Empirical Study of Whether Limiting Lexical Ambiguity Is the Most Effective Way to Improve Machine Translation Accuracy

Hong Xu
Shanghai International Studies University
donnaxu@shisu.edu.cn

ABSTRACT

Baker, K., Franz A., Mitamura, T., and Nyberg, E. (1994) put forward the idea that "the single most useful way to improve the accuracy of a machine translation system is to limit lexical ambiguity." This paper is intended to test whether the argument holds water. The study examined a technical document from four different aspects, and the four versions were translated using Google Translate to answer the question at different levels. The sub-questions concern the impact of controlled language (CL) rules on neural machine translation (NMT) systems, the relationship between lexical constraints and grammatical controls, and the need to discuss CL at the segment and syntactic levels in an NMT environment. The accuracy of the NMT outputs was evaluated by human annotators and BLEU, METEOR, and TER. Consistent with many previous studies, the application of controlled language rules did not seem to affect the NMT quality. But it may negatively impact machine translation accuracy. In addition, limiting lexical ambiguity did not make NMT outputs more accurate than imposing grammatical constraints. Given these two findings, it seems that there is no need to compare lexical and grammatical controls in an NMT environment.

Keywords: Lexical control, Grammatical constraints, Accuracy, Neural Machine Translation, Controlled language.

1. Introduction

To improve the accuracy of machine translation (MT), researchers have turned to controlled languages (CLs) (O'Brien, 2003; Aikawa et al., 2007; Temnikova, 2010; Crabbe, 2017; Marzouk & Hansen-Schirra, 2019; Hiraoka & Yamada, 2019). The general goals of controlled language (CL) are to achieve consistent authoring of source texts and to encourage clear and direct writing. CL is also used to improve the quality of translation output. Authoring with short, concise, and unambiguous sentences increases the chance of achieving high-quality translation (Bernth, 2006). A key element in controlling a source language is restricting vocabulary size and meaning for a particular application domain. Baker, K., Franz A., Mitamura, T., and Nyberg, E. (1994) point out that "the single most useful way to improve the accuracy of a machine translation system is to limit lexical ambiguity."

The present study was mainly designed to test whether the argument proposed by Baker, K., Franz A., Mitamura, T., and Nyberg, E. (1994) holds water. To establish a high-quality research project, an explanation of the controlled languages used in the study is given in Section 2. Section 3 illustrates the reason for choosing neural machine translation (NMT) as the MT tool in the study. Section 4 provides a detailed exploration of the dataset. Section 5 presents the methodologies applied. Section 6 gives examples of CL rules. Section 7 outlines the results. Section 8 presents the conclusions.

2. Controlled Language

A controlled language is "an explicitly defined restriction of a natural language that specifies constraints on the lexicon, grammar, and style" (Nyberg et al., 2003). The notion of CL can be traced back to the work of Charles K. Ogden in the 1930s and 1940s. His "Basic
English” consists of 850 words and a few rules describing how to inflect these words and derive other words.\(^1\)

Mitamura and Nyberg (1995) divide the CL rules into the following categories: lexical, grammatical (sentence and phrase level), and structural controls (text level). Adriaens (1995) adds another type to this list, i.e., Punctuation/Character control.

O’Brien (2003) analyzes eight CL rule sets and selects the most pertinent rules for MT-oriented CLs. Lexical rules involve vocabulary usage, spelling, synonym and pronoun usage, coordination, and others. Syntactic rules are based on the agreement between sentence constituents, repetitions, modifiers, and others. Semantic rules include polysemy. Text structure rules concern sentence length and punctuation. Pragmatic rules mainly refer to textual devices. Syntactic and Lexical rules account for the most significant proportion of rules overall in the group of CLs analyzed.

Considering a significant body of CL rules available, it was essential to identify a set of CL rules suitable for this study. It was decided to adopt the CL rules suggested by Mitamura and Nyberg (1995) for the following reasons. First, the effect of CL rules is associated with the text domains. As both this study and the research conducted by Mitamura and Nyberg (1995) were focused on technical documents, it seemed reasonable to use their taxonomy of CL rules. Second, their classification of CL rules was ideal for the research purpose of this study.

On the one hand, the purpose of the study was to test whether the assumption made by Baker, K., Franz A., Mitamura, T., and Nyberg, E. (1994) confirmed the hypothesis that “the single most useful way to improve the accuracy of a machine translation system is to limit lexical ambiguity”, thus highlighting the control of lexical ambiguity. To do this, we formulated a hypothesis that lexical control is as effective as grammatical control. If this hypothesis were to be confirmed, the assumption made by Baker, K., Franz A., Mitamura, T., and Nyberg, E. (1994) would prove not true. Otherwise, there would be new evidence to support their assumption. How to control the text on the segment and syntactic levels is critical to the study.

On the other hand, Mitamura and Nyberg (1995) claim that CL rules fall into three categories, as mentioned earlier in this section. These three groups include lexical and grammatically controlled documents. This distinction makes it possible to compare the lexically and grammatically controlled documents. Furthermore, the findings may be more persuasive if the approach recommended by Mitamura and Nyberg (1995) is used to test the assumption made by Baker, K., Franz A., Mitamura, T., and Nyberg, E. in 1994.

3. Machine Translation

The translation industry has long been faced with the difficulty of delivering translation services within a limited time and under the pressure of an infinite number of translation projects. Against this background, machine translation tools are increasingly employed to address these issues (Lagarda et al., 2015). There are many machine translation architectures available, such as Rule-based Machine Translation, Statistical Machine Translation, Hybrid Machine Translation, and Neural Machine Translation (Rivera-Trigueros, 2021).

Rivera-Trigueros (2021) claims that neural machine translation is the predominant MT architecture, with Google Translate the most widely applied system. It boosts translation quality (Kalchbrenner & Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2016). It can even reach human parity in some tasks (Hassan et al., 2018). However, Hassan et al. (2018) caution that this result cannot be generalized to all language domains. Klubička et al. (2018) find that NMT tends to sacrifice the completeness of translation to increase overall

---

\(^1\)See [http://ogden.basic-english.org/words.html](http://ogden.basic-english.org/words.html).
fluency. This tendency can result in mistranslation, which may be associated with lexical and grammatical choices. Given these findings, lexical and grammatical controls seem necessary for NMT. Therefore, this paper adopts "Google Translate".

4. Dataset

An operating instruction was chosen for analysis, consisting of 21 sentences. This instruction is entitled 3-Speed Oscillating Stand Fan Operating Instruction and downloaded from Manualzz. The manualzz website (https://manualzz.com) is a universal manual library. It displays manuals for all kinds of products, such as computers and electronics, medical equipment, and pet care. The 21 sentences are translated into Chinese by Google Translate under the four conditions described in Table 1. Thus, the dataset included 84 sentences (21 sentences * 4 versions * 1 translation system). The entire dataset was assessed using automatic evaluation metrics, and professional translators evaluated 22 NMT sentences of the 84 sentences.

Table 1 Four different treatments of the 21 sentences

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DT</td>
<td>The 21 sentences were directly translated by Google Translate.</td>
</tr>
<tr>
<td>2</td>
<td>LT</td>
<td>The 21 sentences were controlled on the lexical level and then translated by Google Translate.</td>
</tr>
<tr>
<td>3</td>
<td>GT</td>
<td>The 21 sentences were controlled on the grammatical level and then rendered by Google Translate.</td>
</tr>
<tr>
<td>4</td>
<td>LGT</td>
<td>The 21 sentences were controlled on lexical and grammatical levels and then translated into Chinese by Google Translate.</td>
</tr>
</tbody>
</table>

The study adopted the taxonomy of CL rules suggested by Mitamura and Nyberg (1995). There are 10 CL rules involved in the study, including five lexical constraints (LC) and five grammatical controls (GC), as specified in Table 2.

The four versions of the 21 sentences were translated by Google Translate. The quality of the four NMT texts was evaluated using automatic evaluation metrics and human assessment. Given the conditions mentioned above, the study aimed to address the following research questions:

Research Question 1

Do lexical and grammatical controls have an impact on NMT quality?

Research Question 2

Do the lexical controls improve the accuracy of NMT more efficiently than syntactic constraints?

2 https://translate.google.com/
Table 2  The CL rules adopted in the study

<table>
<thead>
<tr>
<th>Applied CL rules</th>
<th>Description</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1(LC)</td>
<td>Orthography</td>
<td>By specifying slashes, spelling, capitalization, and hyphenation, see Example 1</td>
</tr>
<tr>
<td>Rule 2(LC)</td>
<td>Encode meaning using synonyms</td>
<td>By replacing words with many meanings with synonyms, see Example 2</td>
</tr>
<tr>
<td>Rule 3(LC)</td>
<td>Encode truly ambiguous terms for interactive disambiguation</td>
<td>By giving ambiguous terms different lexical entries, see Example 3</td>
</tr>
<tr>
<td>Rule 4(LC)</td>
<td>Functional word</td>
<td>By specifying pronouns, see Example 4</td>
</tr>
<tr>
<td>Rule 5(LC)</td>
<td>Participial forms</td>
<td>By replacing participial forms with subordinate clauses, see Example 5</td>
</tr>
<tr>
<td>Rule 6(GC)</td>
<td>Coordination of verb phrases</td>
<td>By avoiding the coordination of verbs or verb phrases, see Example 6</td>
</tr>
<tr>
<td>Rule 7(GC)</td>
<td>Nominal compounding</td>
<td>By avoiding the nominal compounding, see Example 7</td>
</tr>
<tr>
<td>Rule 8(GC)</td>
<td>Verb particles</td>
<td>By replacing verb particles with verbs, see Example 8</td>
</tr>
<tr>
<td>Rule 9(GC)</td>
<td>Clauses introduced by subordinate conjunctions</td>
<td>By ensuring every subordinate clause has a subject and an object, see Example 9</td>
</tr>
<tr>
<td>Rule 10(GC)</td>
<td>Punctuation</td>
<td>By avoiding using ambiguous commas, colons, semicolons and others, see Example 10</td>
</tr>
</tbody>
</table>

This study mainly investigated whether lexical controls are the most helpful way to improve the accuracy of NMT. However, due to the inconsistency among Marzouk (2021), Klubička (2018), and Hiraoke et al. (2019) regarding the impact of CLs on NMT, this study added one more question (Research Question 1) to investigate whether the CL rules affect the quality of NMT. If CL rules do not impact NMT quality, the talk regarding the result of Research Question 2 will not be significant.

5. Methodology

A combination of automatic evaluation metrics and human assessment was applied in the research. The two methods are outlined below.

5.1 Automatic Evaluation Metrics

The study applied a combination of the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2001), the Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Banerjee & Lavie, 2005), and Translation Edit Rate (TER) (Snover et al., 2006) to compare the translation accuracy of the four versions, namely DT, LT, GT, and LGT.

BLEU is used to evaluate the adequacy and fluency of MT outputs by comparing them with reference translations (Papineni et al., 2001). It is designed for general use and can be used for different language pairs (Chatzikoumi, 2020). The MT quality is assessed based on its closeness to the reference translation. The BLEU value for an MT output is rated high if it is close to the reference translation. The score ranges from 0 to 1, with one the maximum (Papineni et al., 2001).

METEOR was built upon the notion of unigram matching between MT output and the reference translation while "capturing how well-ordered the matched words in the machine translation are in relation to the reference" (Banerjee and Lavie, 2005).
TER measures the amount of editing required to make a translation match a reference translation (Snover et al., 2006). The TER score also ranges between 0 and 1. Unlike the other two scores, the higher the TER score, the more editing is needed.

The reference translation in the study was the post-edited version of the 21 sentences and edited by two professional translators with over ten years of experience in the technical domain and English to Chinese translation.

### 5.2 Human Assessment

The human assessment in the study was designed to compare the accuracy of the four different NMT outputs. The accuracy, based on the Multidimensional Quality Metrics (MQM)\(^3\), was classified into four groups: addition, mistranslation, omission, and untranslated.

The human evaluated NMT sentences were reduced from 84 to 21 for the following reasons. First, some sentences in the original texts were not controlled lexically or grammatically. This was because they did not meet the lexical and grammatical control requirements. Second, sentences in different versions overlapped. For example, some in LT are the same as the LGT sentences.

Fiederer & O’Brien (2009) recommend recruiting more than 3 participants. This study involved 5 participants. The annotators were bilinguals, with Chinese as their native language and English as their working language. All were English to Chinese translation practitioners and agreed to join the study. On average, they had worked in the translation industry for more than seven years. They were all familiar with technical documents. They all held a Master's Degree in translation. Each participant was asked to evaluate the entire set of 21 sentences. Participation was remunerated.

Describing the presentation of the various NMT outputs, the participants were first presented with the evaluation guidelines and the background information concerning the text. They were also informed that they would be paid after the assessment. Then they started the evaluation process. They were shown both the source and the corresponding NMT sentences during the evaluation.

Ranking was used as the human evaluation method in the study. Bojar (2011) points out that the approach compares sentences or constituents. It works better than the adequacy and fluency scale (Bojar et al., 2016). Görög (2014) suggests that the sentence to be evaluated should not exceed three. However, Bojar et al. (2016) recommend considering five sentences once. In this study, participants can determine two or more MT sentences by assigning a value, such as 1, 2, 3, or 4, to each sentence.

### 6. Examples of How to Apply CL Rules

This section describes how the source text (ST) is controlled at the lexical and grammatical levels.

#### 6.1 Lexical Control

This part covers five examples of how to impose lexical controls, as specified below.

**Example 1**

ST: To make/stop the fan head oscillate, push down/pull up the oscillating.

LT: To make or stop the fan head oscillate, push down or pull up the oscillating.

---

The slash "/" is replaced with "or".

**Example 2**

ST: Caution
Read Rules for Safe Operation and Instructions Carefully.

LT: Notice
Read Rules for Safe Operation and Instructions Carefully.

"Caution" is replaced with "Notice" not only because "Caution" here and "Warning" in sentence three are all translated as "警告" in Chinese but also because the meaning of "caution" is close to that of "warning". Therefore, "Notice" substitutes for "Caution" to avoid ambiguity.

**Example 3**

ST: Disconnect fan when moving from one location to another.

LT: Unplug fan when a fan is moved from one location to another.

"Disconnect" has more than one meaning, such as "to separate something from something (else)" and "to remove a piece of equipment from a supply of gas, water or electricity." To avoid the amount of ambiguity, "unplug", the synonym of "disconnect", is used.

**Example 4**

ST: Don't let them use the appliances without supervision.

LT: Don't let children and infirm persons use the appliances without supervision.

The use of pronouns is not recommended in controlled language because sometimes it is difficult to figure out what the pronoun refers to. Therefore, "them" is replaced by "children and infirm persons."

**Example 5**

ST: Disconnect fan when moving from one location to another.

LT: Unplug fan when a fan is moved from one location to another.

"Moving" in ST serves as a present participle. However, the use of participle should be restricted. Therefore, the subordinate clause led by when is supplemented with a subject and an auxiliary verb.

**6.2 Grammatical Control**

The following are five illustrations of how to control the sentences at the grammatical level.

**Example 6**

ST: Unscrew and take off the height adjustment ring.

GT: Unscrew the height adjustment ring and remove the height adjustment ring.

Conjoining the arguments of verbs is not recommended. It is advisable to repeat the argument.

**Example 7**

ST: Ensure that the fan is switched off from the supply mains before removing the guard.

GT: Ensure that the fan is disconnected from the power before you remove the guard.
Noun compounding, i.e., a noun-noun structure, is not allowed. Instead, it should be reduced to a single noun. Hence, "supply mains" is replaced by "power".

**Example 8**

ST: Ensure that the fan is switched off from the supply mains before removing the guard.
GT: Ensure that the fan is disconnected from the power before you remove the guard.

Verb particles, which refer to the structures like verb plus preposition, and verb plus adverb, are also not a good option. They are best replaced by a single verb. Here, "disconnect" is used rather than "switch off". The reason why "disconnect" fits here rather than in Example 3 is that the two meanings embedded in the word "disconnect" work well in this context without causing any possible confusion.

**Example 9**

ST: Ensure that the fan is switched off from the supply mains before removing the guard.
GT: Ensure that the fan is disconnected from the power before you remove the guard.

Each complex sentence should contain a subject and an object. So, the preposition phrase "before removing the guard" is replaced with "before you remove the guard."

**Example 10**

ST: DO NOT use fan in window, rain may create electrical hazard.
GT: DO NOT use fan in window. Rain may create electrical hazard.

The comma here is unacceptable. Therefore, the comma is replaced with a period.

7. Results

7.1 Automatic Evaluation Results

There are three dependent variables in the automatic evaluation tests: BLEU, METEOR, and TER, and one categorical variable at four levels, namely LGT, DT, LT, and GT. The Shapiro test is used to check for normal distribution. The results show that the dependent variables, such as BLEU (p = 0.9719), METEOR (p = 0.1612), and TER (p = 0.8822), are all normally distributed.

T-test suggests that LGT and DT are not statistically different regarding automatic evaluation metrics (p = 0.8202). This result is consistent with the research conducted by Marzouk (2021) and Klubicka (2018). However, there is a slight difference (0.01) between LGT and DT in terms of TER (Table 4). Similarly, the difference between LT and GT is not statistically significant (p = 0.8516). Nonetheless, the TER of LT is slightly lower than that of GT (0.05).

There is also no significant difference between groups like LGT and LT (p = 0.4864), LGT and GT (p = 0.4746), DT and LT (p = 0.7735), as well as DT and GT (p = 0.5979).

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>METEOR</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGT</td>
<td>0.44</td>
<td>0.67</td>
<td>0.3</td>
</tr>
<tr>
<td>DT</td>
<td>0.45</td>
<td>0.67</td>
<td>0.39</td>
</tr>
<tr>
<td>LT</td>
<td>0.46</td>
<td>0.64</td>
<td>0.43</td>
</tr>
<tr>
<td>GT</td>
<td>0.47</td>
<td>0.63</td>
<td>0.48</td>
</tr>
</tbody>
</table>

**Table 4** The BLEU, METEOR, and TER values of DT, LT, GT and LGT
Human Evaluation Results

In the human evaluation tests, all the data were tested with the Shapiro test. The results showed that LT (p = 0.3006), DT (p = 0.5536), LGT (p = 0.2985), and GT (p = 0.4634) were normally distributed.

Two-tailed t-tests show a significant difference between LGT and DT (p = 0.0158), which contradicts the finding based on the automatic evaluation. However, there is no significant difference between GT and LT (p = 0.2736).

Two-tailed tests were also carried out among other groups, such as LGT and LT, LGT and GT, DT and LT, and DT and GT. There was no significant difference among LGT and LT (p = 0.06163), LGT and GT (p = 0.6061). However, there were significant differences among DT and LT (p = 0.01669), and DT and GT (p = 0.01585). Wherever DT was involved, there were statistical differences. This indicates that the annotators all believed that DT was superior to the controlled texts in terms of accuracy.

8. Conclusion

The study was mainly intended to determine whether there is a significant difference between lexical controls and grammatical constraints in an NMT environment. This had not been investigated in previous research to the best of my knowledge. It was also hoped that the study would shed light on the impact of CL on NMT.

Consistent with the findings from some of the previous research, this study also suggests that CL rules may not improve the output of neural NMT. Another result worth noting in this study is that the application of CL rules may even downgrade the NMT quality. This could be associated with Google Translate. The Google NMT is built on natural languages in real scenarios and not controlled languages. Furthermore, this study did not consider the specific characteristics of NMT systems. The NMT system is different from other MT architectures, such as Statistical Machine Translation and Rule-based Machine Translation. NMT has posed new challenges for translation. Therefore, the CL rules should take these new challenges into account.

Most importantly, the assumption made by Baker, K., Franz A., Mitamura, T., and Nyberg, E. (1994) proves to be not true in this study. The automatic and human evaluation results demonstrate no significant difference between the LT and GT in terms of NMT accuracy. Given that CL rules may not boost the NMT quality, the discussion about whether the lexical restrictions approach is more effective than grammatical ones in NMT systems may well sound meaningless. However, whether the finding of this study can be generalized still needs further investigation as this study only involved the translation of a simple operation instruction from English to Chinese.

Finally, it should be noted that the findings of the automatic evaluation were slightly different from those of human assessment. This may be attributed to the reference translation. Reference translations are regarded as the gold standard for automatic evaluation, but some translation outputs are considered flawed simply because they are not close to the reference translation (Chatzikoumi, 2020). During the human assessment, the reference translation was not adopted as the study asked participants to rank translation outputs from best to worst. This discrepancy makes it necessary to introduce innovative technologies to measure NMT quality, such as eye-tracking and key-logging.
References


Who needs an MQM Scorecard?

Catherine Marshall Martins  
*Brigham Young University (BYU)*  
catherine.marshall@outlook.com

Alan K Melby  
*Vice-President (FIT)*  
Emeritus Professor, BYU  
President, LTAC Global  
alan.melby@fit-if.org

ABSTRACT
The presentation and this paper describe reference software called the MQM Scorecard. This software is free and open source. It runs in a browser, and organizations desiring to use it must install a copy on their own Web server. Then, they can set up Translation Quality Evaluation (TQE) projects and assign them to evaluators. The results of a TQE project are available as a JSON file for machine processing. The JSON file can be converted to an HTML file for human consumption or searched using software for analysis.

The first version of MQM (Multidimensional Quality Metrics) was developed several years ago within the European Commission’s QT21 project. MQM is the basis of a forthcoming international standard.

MQM is intended for manual, reference-free evaluation of either human or machine translation.

This presentation not only describes the MQM Scorecard but also provides some context on three related TQE standards that are currently under development. This paper serves as a practical guide on the use of the MQM Scorecard, to help someone who wants to conduct an analytic TQE decide whether the MQM Scorecard might be a helpful tool to accomplish this task.

So, who needs an MQM Scorecard? Anyone who wants to use an MQM approach to measuring the quality of either human translation, post-edited machine translation, or raw (unedited) machine translation.

1. Introduction

Multidimensional Quality Metrics (MQM) is a framework used for classifying errors in either human or machine translations. See the MQM website ([www.theMQM.org/](http://www.theMQM.org/)) for more information. The MQM Scorecard is a free and open-source software app that facilitates translation quality evaluation within the MQM framework.

MQM uses an analytic translation quality evaluation (TQE) methodology. Analytic evaluation involves identifying errors at the word or phrase level. This contrasts with a holistic methodology, which identifies properties of a translation at the macro level, such as an entire document. These two methodologies are complementary. A holistic evaluation is less time-consuming, while an analytic evaluation supports the correction of specific errors in the current translation work product and improvement of the translation process to reduce errors in future translations. Because analytic evaluation requires the expertise of a human bilingual and is time-consuming and thus expensive, sampling techniques are often employed. That is, only selected documents or portions of documents are evaluated analytically. Sometimes a holistic evaluation is performed to decide whether an analytic evaluation is warranted.

Analytic TQE is often conducted using a scorecard. In the translation sector, many scorecards have been and are used. Long ago, some scorecards were on paper. Today many scorecards are Excel spreadsheets. What all scorecards have in common is that they include counts of instances of various types of errors in the translation being evaluated, along with an overall quality score. Typically, each error is assigned one or more penalty points, according to
the severity level of the error. The MQM Scorecard described here is not a spreadsheet. It is a software app that runs partly in a browser and partly on a server. The browser component communicates with a server-side database that is shared among multiple evaluation projects and human evaluators. Each installation of the MQM Scorecard uses one error typology but supports multiple TQE metrics. Before an evaluation project begins, the source and target texts must be segmented and aligned, forming a bitext consisting of a sequence of translation units. Also, an MQM-compliant TQE metric, based on the translation project specifications and compliant with the error typology, must be available. The bitext and the metric are loaded into the MQM Scorecard, which configures itself according to the error types in the metric. Then, a human evaluator identifies errors, each associated with a particular translation unit. The results of the evaluation can be exported to an annotation file that contains a copy of the bitext and the identified errors. The annotation file can then be further processed as needed, for example, to produce reports.

2. **Overview of the Paper**

The main purpose of this paper is to help someone who wants to conduct an analytic TQE decide whether the MQM Scorecard might be a helpful tool to accomplish this task.

The following sections of the paper are based on the slides used in a presentation at the AsLing TC43 conference. Those slides are available on the AsLing website (www.asling.org/tc43/18-November-2021/) and on the Translation, Theory, and Technology website (https://www.ttt.org/wp-content/uploads/2022/03/Melby-Who-needs-an-MQM-Scorecard-2022-03-12.pdf). The slides can be consulted while reading this paper and used as an outline. This is admittedly an unconventional approach, but this is not an academic paper. It is a practical guide preceded by an explanation of how analytic evaluation of translation quality relates to the rest of quality management.

3. **The Conference Presentation** (see Slide 1)

The live presentation (Who needs an MQM Scorecard?) during the AsLing TC43 conference was made on November 18, 2021, by Catherine Marshall Martins and Alan Melby. The conference was entirely virtual.

4. **About MQM** (see Slide 2)

As explained in the Introduction section of this paper, the MQM Scorecard is tied to MQM. In MQM, there is only one error typology, but there are multiple custom metrics based on it. Each metric is tied to a set of translation specifications. A set of translation specifications is based, in turn, on the standard set of translation parameters found in ASTM F2575-14 Section 8 and summarized in Annex B of ISO 17100. Information on translation parameters is also available on the Tranquality website (www.tranquality.info).

Among the standard translation parameters is the question of which termbase to use. A terminology error in an MQM evaluation is, of course, relative to the specified termbase.

Another translation parameter is the question of which style guide is authoritative. Style errors will then be relative to the specified style guide.

A less mechanical translation parameter is the purpose of the translation. If a translation does not fulfil its purpose, that failure needs to be expressed in terms of the MQM error typology.

Overall, MQM errors indicate deviations from the holistic expectation that the target text is fluent and corresponds to the source text, relative to the purpose, intended audience, and other specifications of the translation project.
Thus, an MQM-based evaluation of the quality of a translation uses the ISO 9000 approach to measuring quality, namely, how well the translation work product meets the project specifications. Sometimes, possible translation issues (such as apparent discrepancies between a translation and a termbase) are detected by an automatic quality assessment tool. However, just as with spell checkers, some errors are incorrectly flagged or not detected. Deviations from the specifications are detected or confirmed by a human evaluator and annotated using the MQM error typology. The connection between error detection and quality management (QM) is discussed in the next section.

5. **MQM vs Reference-based Metrics; Quality Management** *(see Slide 3)*

MQM evaluations are manual and reference-free. Thus, they are unlike reference-based metrics such as BLEU or METEOR in which a computer automatically compares a translation against a reference translation that was previously created by a human translator. Since MQM does not require a reference translation, an MQM-based evaluation cannot be conducted in a fully automatic fashion. It requires the participation of a bilingual human who understands the specifications of a project and thus the criteria on which a translation is being evaluated.

Reference-based and reference-free evaluation methods should not be viewed as competing with each other. Reference-based methods are typically used when fine tuning a machine translation engine applied repeatedly to the same source text in a research environment. Reference-free methods, on the other hand, are typically used when evaluating a translation in a commercial environment.

A language service provider might be revising a translation to determine whether it is ready to be delivered to the requester. Or the requester might be evaluating a translation to decide whether to accept it. Obviously, no reference translation is available in such environments. If the revision detects errors, these are typically corrected. In evaluation, it depends. Sometimes, errors detected in the evaluation are corrected. Sometimes, they are left uncorrected.

In both cases, detection and correction are associated with aspects of QM when applied to translation projects. The five components of QM treated within MQM are planning, control, improvement, assurance, and evaluation.

Quality Planning (QP) in the translation sector involves setting up a translation process that is expected to produce quality results, that is, translations that meet agreed-on specifications.

Quality Control (QC) involves detecting errors in real-time, during or between the various steps involved in producing translations. Correcting those errors, while an important part of translation project management, is not always central to QC, which is focused on real-time adjustments to the process used rather than fixing a product. This raises the question of where correction of translation errors fits into quality management as it relates to translation projects.

Producing translations is different from producing inexpensive, mass-produced widgets. A cheap widget that fails quality control is sometimes simply discarded. If multiple widgets are defective in the same way, the Quality Control manager might pause production briefly to look for a defect in the production process that can be corrected quickly so that production can continue more effectively.

Substantial translations are more like projects, such as building houses. Minor flaws in a house detected during an inspection, which is a form of quality control, are typically corrected. Only major flaws, such as an incorrectly poured foundation or the use of inadequate building materials, result in tearing down a house and starting over.
Correcting translation errors can actually be associated with several aspects of quality management. Correcting errors so that a translation product passes inspection can be part of either QC or Quality Evaluation (QE), depending on the QM system. Analyzing errors in order to improve the process that will be used in future translations is usually conducted under Quality Improvement (QI).

Although the primary use of MQM is for evaluation, it can be used to assess a translation when providing feedback to a language professional, such as the translator or the reviser in a translation project. In that case, the application of MQM might stop short of evaluation, which is a judgment of whether a work product passes or fails. Instead, the individual errors that were identified by the evaluator are discussed in a constructive learning environment.

Helping a translator improve their skills, based on errors detected in their translation, is part of QI, rather than QE. Instead of simply annotating errors in a particular translation work product, QI is focused on the translation process. The question then becomes how to avoid similar errors in future translations. Improving the competence of a translator can reduce the number and severity of errors in future translation projects. Other aspects of translation process improvement might include maintenance of translation memory databases and terminology databases or retraining of a machine translation system that provides translations of segments. The translation process also involves the selection of human resources for a particular project and communication between the requester and the provider. Ideally, applying an MQM metric includes identifying the root causes of translation errors and contributes to an improvement of the translation process.

Quality Improvement feeds into Quality Assurance (QA), which involves an audit of the entire translation system, including human and data resources. Often, the results of a QA audit are used to improve the alignment between a particular translation system and the needs of a major customer or group of customers. The desired result of a QA audit is to give stakeholders assurance that the translation system can produce quality translations. Or indicate to the management team how to re-align the process so that the next translation will go better. QA cannot ensure, that is guarantee, that every translation produced in the future will meet specifications. However, it is intended to help stakeholders be more confident that a translation system can produce quality translations. MQM is important to quality management in that it can, when used by competent human evaluators, inform both QC (conducted in real-time during production) and QE (typically conducted after production) by identifying deviations from agreed-on specifications, that is, translation errors. With QC or QE, detected errors can be corrected so that the product is judged to be acceptable. In the QI-QA cycle, errors detected during QC or QE can be used to improve the process that will be applied to future translations.

An important point is that automatic, reference-based TQE methods, such as BLEU, are not analytic. A higher BLEU score indicates, to some degree, that one translation is closer to the reference translation than another. But BLEU scores are holistic. They do not indicate why one translation is better than another or how to improve the translation system. Thus, BLEU cannot be part of a quality management system.

MQM, on the other hand, can be an important component of a translation quality management system, and the MQM Scorecard can help evaluators with their work.

The application of MQM to commercial translation systems involving human bilinguals is obvious, especially when the initial translation is produced by a professional human translator. However, the application of MQM to environments where raw NMT (neural machine translation) is sent directly to the consumer is less obvious, but it is being explored. There is no straightforward way to tell an NMT system to avoid making errors of a certain type found in the MQM typology. Some would say this because current AI-based MT systems do
not really understand what they are translating, but the question of understanding goes far beyond the scope of this paper.

A very different kind of reference-based evaluation is found in some translator certification programs. For example, the ATA (www.atanet.org) certification program involves multiple candidates translating the same source text and multiple human evaluations of each translation. When a new source text is chosen, the certification committee produces a reference translation, but there is no automatic comparison between a candidate’s translation and the reference translation. Indeed, the reference translation is only consulted as needed.

6. Related TQE standards (see Slide 4)

The 2015 version of MQM was developed as part of the QT21 project funded by the European Union’s Horizon 2020 research and innovation program under grant agreement No. 645452 (www.qt21.eu/quality-metrics/). Under this project, the error typologies of the initial version of MQM and the initial version of TAUS Dynamic Quality Framework (DQF) were harmonized so that the DQF typology became a strict subset of the MQM typology.

The developers of the 2015 version of MQM realized that once the QT21 project had ended, the participants would turn to other projects. Long-term maintenance of MQM would then become a concern. It was decided that ASTM International would be the only organization licensed to use the trademark MQM and develop MQM into an international standard. An ASTM project was initiated in 2014 (www.astm.org/workitem-wk46396) and, although it was somewhat dormant in the beginning, has become quite active in recent years. It is anticipated that the draft MQM standard will go to ballot in 2022.

The first version of MQM had over one hundred error types which were categorized within seven overarching dimensions. The current version of MQM is focused on the first two layers of the original error typology (themqm.org/error-types/). There have been several minor changes in the second layer since 2015, but the top layer, the dimensions, remains essentially unchanged except for two name changes.

The dimensions are terminology, accuracy, linguistic conventions (called fluency in MQM version 1), style, locale conventions (which ensures that all information is displayed properly for the locale, such as dates or times), audience appropriateness (called verity in MQM version 1), and design and markup. The dimensions and the second level, consisting of subtypes of each dimension, form the MQM Core typology.

ASTM and ISO (www.iso.org) are two of the largest and oldest standards bodies in the world. Over the past decade, there have been several attempts within ISO to develop a standard that includes translation quality evaluation. The first attempt, ISO 14080 ("Assessment of translations") was terminated during the June 2012 meeting of ISO TC 37 in Madrid.

The second attempt, ISO 21999, was terminated at the June 2019 meeting of ISO TC 37 in Ottawa.

The third attempt, ISO 5060, is under active development. Unlike the two previous projects, the 5060 project is being carried out in collaboration with the ASTM MQM project. Thanks to the work of a joint task force, the error typology in ISO 5060 will likely be compatible with the MQM Core typology. This would be good news for the translation sector and professional translators, because it would facilitate communication about translation errors across the sector and the profession.

There is also an active project within ASTM to develop a holistic translation quality evaluation standard (WK54884). Ongoing coordination between the analytic and holistic projects within ASTM has prevented a naming conflict regarding error types. Within the draft
holistic standard, there are only two error types: correspondence and fluency. Neither of these error types matches exactly any of the MQM dimensions. Correspondence includes the MQM dimensions Terminology and Accuracy, while Fluency includes the MQM dimensions linguistic conventions and style. That is why it was agreed between the two projects not to use any of the seven MQM dimensions to name either of the two holistic evaluation categories.

7. Needed to set up an MQM Scorecard project (see Slide 5)

We now turn to the operational details of the MQM Scorecard.

In order to set up a project in the MQM Scorecard, it is necessary to upload an MQM Typology file in the form of an XML file. This file is the entire set of error types that the projects on a Scorecard instance will be using. The error typology file is uploaded once into the Scorecard and can be accessed and downloaded for later reference.

Another requirement is a Metric file, also an XML file, which is a subset of the Typology. This file includes any specifications relevant to a project (or a group of projects). This will be uploaded with the creation of each Scorecard project, and therefore can differ from project to project. However, all the errors defined in the Metric file must also be included in the Typology file for the Scorecard to function properly.

The last things needed to set up an MQM Scorecard project are a bitext file and an optional specifications file. The bitext file is a plain UTF-8 text file where each line is a single segment in the source text followed by a tab and the corresponding target segment. This file can start from an ordered translation memory (TMX file) or an XLIFF which is then converted to a TXT file. The specifications file is an STS file which contains the specifications relevant to the project. While the specifications themselves are not optional, it isn’t necessary to upload them to the Scorecard, but it can be useful to reference later. The translation specifications can exist outside the Scorecard if needed.

8. Current Projects using the MQM Scorecard (see Slide 6)

The MQM Scorecard is currently being used by two TQE projects:
- The ATA Database project and
- The Whale project.

The ATA Database project is a joint effort of BYU (www.byu.edu) and KSU (www.kent.edu). The project includes taking ATA Translation Certification Examinations from previous years and entering them into a database in a machine-processable format for future translation students, teachers, and institutions to reference.

For an exam to end up in the database it needs to pass through the following steps. First, scanned PDF documents of the exam need to be transcribed to two separate text files, one containing the source text and the other containing the target text. Then the two text files are entered into a Computer-Assisted Translation (CAT) tool (currently MemoQ is being used) where the text is aligned and exported to a TMX file that retains the sequence of translation units in the source and target texts. That TMX file is then processed by a Python script that creates a bitext file in the form of a TXT file.

The MQM Scorecard comes into play once the bitext file is created. An MQM Scorecard project is created using that bitext file and a metric XML file which is the subset of the MQM typology. The metric was created based on the needs of the ATA examinations. For example, since the ATA examination translations were handwritten and later transcribed to plain text files, there was no concern for the design and markup of the translation file. Thus, the metric used for those projects did not include that MQM dimension.
Once an MQM Scorecard project is created, the grader’s annotations of errors on the ATA exam are carefully entered into the Scorecard using the MQM framework of errors. All information on the exam is included, such as the marked errors, the severity level of each error, and any additional comments left by the grader. ATA grading framework codes are placed in the notes.

Once all the errors noted by the grader have been entered into the Scorecard project, the project is saved and marked as completed. An option is given to download a JSON file which contains all the information added into the Scorecard project. This file allows the data to be used in any way that the creator of the Scorecard project may want. For the ATA Database Project, it is then used to extract any information that needs to be included in the Scorecard, such as the source and target texts, the grader’s annotations, comments, etc. This information will then be displayed in the database in an HTML file.

The Whale project falls under the FIT (www.fit-ift.org) translation quality evaluation task force. The name derives from the fact that it involves multiple independent evaluations of a translation that involves whales. The objective is to develop an evaluation system, including professional human translators, evaluator training materials, and the MQM Scorecard, that results in evaluations with an acceptable level of IRR (Inter-Rater Reliability). This is important because in the past, many manual, reference-free translation quality evaluation systems have displayed uncomfortably low levels of IRR.

9. Running the MQM Scorecard

The MQM Scorecard is a browser-based program and therefore does not require any downloading of programs or installation on a client computer. Each instance of the Scorecard is self-contained.

The MQM Scorecard was originally created in 2014 by Tyler Snow, an MA student at BYU. It was in PHP and later ported to Symfony by the QT21 project. In 2021, the Scorecard was ported to a JavaScript framework. This porting to the new framework was funded by BYU and LTAC.

The MQM Scorecard is applicable for anyone who wants to produce an analytic, reference-free evaluation of a translation within the MQM framework. It can be used for a variety of projects since it allows customized MQM TQE metrics, based on a set of translation specifications.

10. BYOS: Bring your own server

The MQM Scorecard runs on a Linux server and protects any data that is contained in the Scorecard. Thus, privacy is not an issue with the MQM Scorecard approach, as it is with any TQE system that stores translations and evaluations on a server not controlled by the organization conducting the evaluations.

The server requirements include root access to the server and the ability to run React and Express. The software requirements include Git, PostgreSQL v9.x, and Node.js v16.x. The hardware requirements include a 1 GHz processor, 2 GB of RAM, and 512 MB of hard disk drive.

11. Demonstration

The next few sections review the demonstration part of the presentation. Adding a typology, creating a Scorecard project, and exporting the information to a JSON file are all covered here.
12. MQM Error Typology (see Slide 10)

To begin to create projects in the MQM Scorecard, the first requirement is to input the MQM Error Typology. The typology will only be input once into the MQM Scorecard. It does not need to be the complete MQM typology, but every metric file (the subset of the Scorecard) needs to be included in the typology. This is an XML file that is input in the Manage Typology tab in the Scorecard.

13. Creating a Project (see Slide 11)

The creation of a Scorecard project is done in the Create Project tab in the Scorecard. This is where the creator of the project will add all the required parts to create a project. This includes the bitext file, specifications file, and metric file. A project name can also be added here, but it can be edited later. Hit Submit to create the project.

14. Scorecard Interface (see Slide 12)

The project can be viewed in the View Projects tab. Here in the Scorecard interface, the source text is found on the left side of the table and the corresponding target text is found on the right side. The notes section is found to the right of the table.

15. Adding an Issue (see Slide 13)

To add an error to the text, the evaluator will click on the segment where they want to add the error. It will be outlined in red once it is selected. Then they will click on the highlight icon on the right side of the target text. Once that is selected, they can highlight the text that contains the error they want to mark. Then a pop-up will appear where they can select the MQM error type and the severity level. The error types listed will be based on the ones chosen in the metric file that they uploaded when creating a new project.

16. Notes Section (see Slide 14)

After the evaluator selects the MQM error and the severity level, they can hit Continue and a new pop-up will open that will allow them to add additional notes to the error. Once they have done that, they will click on the Add New Error button to create the error.

17. Seeing Errors Using the Interface (see Slide 15)

After creating the error, the evaluator can see it in the Scorecard interface. By clicking on the name of the error, they can see the text that was highlighted and the notes that they added. There is no way to edit an error once it has been created. If any changes need to be made to the error, the evaluator will need to delete the error and create a new one with the appropriate changes.

18. Error Summary (see Slide 16)

The Reports tab within the project will open a spreadsheet that contains statistical information about the project. It shows the number of errors for each error type and their corresponding severity level. At the bottom of that screen is a button to export all the project information as a JSON file.

19. JSON File Information (see Slide 17 and Slide 18)

A detailed description of the MQM JSON format can be found on the Translation, Theory, and Technology website (https://www.ttt.org/mqm-json/).

20. Conclusion (see Slide 19)

The answer to the question of who needs an MQM Scorecard is someone who is conducting analytic translation quality evaluations in the context of quality management. They
are using a reference-free evaluation method, meaning that their evaluators understand the specifications of the translation and will judge the translation against those specifications. To use this MQM Scorecard, the parent organization will need to have their own or a rented Linux server. There are other, more sophisticated MQM-compatible TQE tools available commercially, but this one is completely free and can be adapted to a number of possible projects.

**Biographies:**

**Alan Melby** was raised in Indiana (USA). He still identifies as a Hoosier (https://en.wikipedia.org/wiki/Hoosier). He became fascinated with translation in the mid-1960s, while on study abroad in St Brieuc, France, as a high school student. Joined a machine translation project in 1970 while working on a Bachelor of Science in mathematics. By 1978, after obtaining a PhD in computational linguistics, experienced an intellectual crisis regarding the nature of language while an assistant professor of linguistics at Brigham Young University, concluding that unambiguous general language would be the ultimate prison but domain-specific language can and should be unambiguous. In 1979, he shifted focus toward tools for human translators. In the 1980s, he studied translation and became an ATA-certified French-to-English translator. In the 1990s, he developed an interest in the philosophy of language and wrote a book about human and machine translation (*The possibility of Language*) with a philosopher, Terry Warner.

In the 21st century, he has focused on service to the translation profession, currently (2021) serving as vice-president of the International Federation of Translators (FIT), and collaborating on the development of translation-related standards, especially TBX (ISO 30042), for the exchange of information between terminological databases. In 2014, he retired from full-time teaching as a full professor and became an emeritus professor and the president of a small non-profit (LTAC Global). Since 2015, standards work has expanded to translation quality evaluation within the MQM framework (Multidimensional Quality Metrics for Translation Quality Evaluation) under the umbrella of ASTM International (www.astm.org), a standards body.

If you wish to install a copy of the MQM Scorecard on your Linux server, please contact me at: alan.melby@fit-ift.org (or my private email address, melbyak@yahoo.com) with the keywords “MQM Scorecard” in the subject line.

**Catherine Marshall** was raised in Michigan (USA). She is a student in the Master’s program in Linguistics at Brigham Young University (Provo, Utah, USA campus). She graduated from Brigham Young University in April 2021 with a Bachelor’s in Portuguese and Translation/Localization. She has worked on the ATA Database project since February 2021, where she oversees all data entry, including the annotation of ATA (www.atanet.org) certification exam errors using the MQM Scorecard.
MT Quality and its effects on post-editors and end-users

Maria Stasimioti  
Department of Foreign Languages, Translation and Interpreting  
Ionian University  
stasimioti@ionio.gr

Vilelmini Sosoni  
Department of Foreign Languages, Translation and Interpreting  
Ionian University  
sosoni@ionio.gr

Abstract
The study investigated whether the quality of the raw machine translation (MT) output generated by a statistical machine translation (SMT) system, on the one hand, and a neural machine translation (NMT) system, on the other, affects not only the temporal, technical and cognitive effort expended by translators while performing full post-editing (PE) of the MT output, but also the extrinsic quality of the final post-edited text, i.e., its acceptability by the end-users in terms of readability and comprehensibility (Gouadec, 2010; Suojanen et al., 2015). The investigation was based on: (i) automatic and human evaluation metrics for the quality of the MT outputs, (ii) eye-tracking and keystroke logging data to measure the temporal, technical and cognitive effort during PE and (iii) a reception study based on a Likert-type scale to measure reader satisfaction. The findings indicate that the lower quality of the SMT output did affect the actual PE effort, increasing the temporal, technical and cognitive effort expended by the translators, but it did not lead to translations of inferior quality according to the final end-user.

1 Introduction and related work

In recent years, machine translation (MT) quality has improved considerably mainly as a result of the development of neural machine translation (NMT) models. MT has thus been increasingly used in industrial settings to produce raw translations to be further post-edited by translators (Lommel and DePalma, 2016; Koponen, 2016).

Post-editing (PE) is influenced by the MT system used, as well as the number and the type of errors, given that some errors have been found to be more demanding than others (Koponen, 2012). Numerous studies have shown that the improved quality of the NMT system output, especially at the level of fluency, requires the correction of fewer segments, mainly due to the lower number of morphological errors (Castilho et al., 2017a, 2017b). However, this does not always result in lower PE effort, mainly due to the fact that NMT errors at the level of adequacy are more difficult to identify and correct than the obvious word-order errors and disfluencies occurring in phrase-based machine translation (PBMT) and statistical machine translation (SMT) outputs (Castilho et al., 2017b), requiring longer post-editing times (Carl and Báez, 2019).

The differences between various MT systems, as regards the quality of their output and the types of errors, are reported by several recent studies. Some (Bahdanau et al., 2015; Jean et al., 2015; Junczys-Dowmunt, 2016; Dowling et al., 2018; Deep et al., 2021) relied on automatic evaluation metrics (AEMs) like BLEU (Papineni et al., 2002) and HTER (Snover et al., 2006); others used human evaluations of the MT output quality, employing adequacy and fluency ratings (Bentivogli et al., 2016), manual error analyses (Klubička et al., 2017, 2018; Popović, 2018) or a combination of methods (Burchardt et al., 2017; Castilho et al., 2017a, 2017b, 2018; Toral and Sánchez-Cartagena, 2017; Esperança-Rodier et al., 2017; Shterionov et al., 2018; Mahata et al., 2018; Koponen et al., 2019; Jia et al., 2019; Mual et al. 2019; Esperança-Rodier and Rossi, 2019; Stasimioti and Sosoni, 2019, 2020).
Although automatic and human evaluation metrics are very important for the evaluation of the quality of the final product, they do not provide any indication about the acceptability of the final product by its end users, i.e., its readers. For this reason, reception studies are needed. Reception studies focus on the audience, their comprehension, appreciation or rejection of what reaches them through the medium of translation. Although these are crucial since they offer “a uniquely thorough picture of the life and afterlife of these texts” (Di Giovanni and Gambier, 2018: n.p.), so far they have received limited attention in Translation Studies (TS), even though reception study theories have been applied in recent decades, first to literary translations and then to other areas and text types. As Gambier (2018: 44) aptly observes, this failure to view the reader as the re-interpreter of the translators’ work in TS has obstructed the study of reader response and has also “risked placing the translator as the ultimate controller of textual meaning, at least for the target culture audience, and reinforced the transmissionist model by assuming that the translator’s interpretation reaches the reader intact” (ibid). Interestingly, while the study of translation and its reception is limited, in the case of PE there are several studies that focus not only on measuring the relative quality of post-edited texts using automatic and human evaluations metrics, but also on the readers’ views and engagement. In particular, Castilho, O’Brien, Alves, and O’Brien (2014), in a study measuring the usability of raw machine translated output and post-edited output for instructional text relating to a commercial PC security product, found that PE significantly increases the usability of machine-translated text. More recently, Hu, O’Brien and Kenny (2019) investigated Chinese end users’ reception of massive open on-line course (MOOC) subtitles that had been translated from English into Chinese under three conditions: human translation, raw MT, i.e. unedited, and post-edited MT. They found that participants who were offered full post-edited subtitles scored slightly better on the reception metrics than those who were offered raw MT subtitles, while surprisingly those who were offered human translation did not perform better than the other two groups.

The aim of the present study is twofold: to investigate whether the quality of the raw MT output generated by an SMT system, on the one hand, and an NMT system, on the other, affects not only the temporal, technical and cognitive effort expended by translators while carrying out full PE of the MT output, but also the extrinsic quality of the final post-edited texts, i.e. its acceptability by the end-users in terms of readability and comprehensibility (Gouadec, 2010; Suojanen et al., 2015).

2 Methodology

The investigation is based on: (i) automatic metrics, i.e., BLEU, METEOR, WER and TER, and human evaluation metrics, i.e., side-by-side ranking, adequacy and fluency rating, and error annotation, to evaluate the quality of each MT output, (ii) eye-tracking and keystroke logging data to measure the temporal, technical and cognitive effort expended by ten MA Translation students while post-editing each MT output in Greek, (iii) a reception study based on a Likert-type scale to measure reader satisfaction.

2.1 Participants and training

Ten Greek students enrolled in the Science of Translation MA offered by the Department of Foreign Languages, Translation and Interpreting (DFLTI) of the Ionian University during the 2018-2019 spring semester participated in this study. The participation in the experiments was voluntary and all participants signed a consent form, while all stored data were fully anonymised in accordance with Greek Law 2472/97 (as amended by Laws 3783/2009, 3917/2011 and 4070/2012). As can be seen in Table 1, all participants were female, the majority in the 18-24 and 25-34 age groups, and had an undergraduate degree in Translation or in a related field, such as English Studies or Greek Linguistics. Only three of the participants had professional experience in translation, while none of them had experience in PE.
Table 1: Participants’ gender, age distribution, education level, degree type and experience in translation and PE

<table>
<thead>
<tr>
<th>Gender</th>
<th>Female</th>
<th>Male</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age distribution</th>
<th>18-24</th>
<th>25-34</th>
<th>35-44</th>
<th>45-54</th>
<th>&gt; 55</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education level</th>
<th>Undergraduate degree</th>
<th>Postgraduate degree</th>
<th>PhD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Degree type</th>
<th>Translation</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experience in translation</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experience in PE</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

As familiarization with PE was a prerequisite for participating in this study, 12 hours of training (six two-hour sessions) were offered to the students in the context of the compulsory MA module in Translation Technology to introduce the students to MT and PE as well as to the recent developments in the respective fields. Upon completion of the training, students were expected to be able to (i) use MT during the pre-translation process, (ii) evaluate MT output using both automatic and human evaluation metrics, and (iii) post-edit MT output according to the expected level of quality (full/light PE). The main topics covered included the theory and history of MT and PE, the basic principles of MT technology, analysis of the dominant systems in the market, the importance of controlled language and pre-editing for MT, quality metrics and evaluation of MT output, PE quality levels, PE effort and productivity (temporal, technical, and cognitive effort), MT output error identification, MT engine implementation in the translation workflow, and post-editor profile and associated skills (O’Brien, 2002; Depraetere, 2010; Doherty et al., 2012; Doherty and Kenny, 2014; Kenny and Doherty, 2014; Koponen, 2015; Guerberof and Moorkens, 2019).

2.2 Source texts

The source texts (STs) used in this study were two (2) short (~140 words) semi-specialised texts about the 2019 EU elections selected from the British daily newspaper The Guardian. These were informative, journalistic texts aimed at the educated layman (Newmark, 1988: 13). More specifically, both texts had comparable readability scores (between 1200L and 1300L), i.e., they were suitable for 11th/12th graders. The Lexile Analyzer was used, as it relies on an algorithm to evaluate the reading demand—or text complexity—of books, articles, and other materials. In particular, it measures the complexity of the text by breaking down the entire piece and studying its characteristics, such as sentence length and word frequency, which represent the syntactic and semantic challenges that the text presents to a reader. The outcome

---

4 [https://lexile.com](https://lexile.com)
is the text complexity, expressed as a Lexile measure, along with information on the word count, mean sentence length, and mean log frequency.

<table>
<thead>
<tr>
<th>Lexile® Measure</th>
<th>Text 1</th>
<th>Text 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sentences</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Mean sentence length</td>
<td>23.50</td>
<td>20.00</td>
</tr>
<tr>
<td>Word count</td>
<td>141</td>
<td>140</td>
</tr>
</tbody>
</table>

Table 2. Lexile® scores for the STs used in the study

2.3 MT outputs

Each ST was machine-translated without any pre-editing using the SMT system developed by Google and the NMT system developed by Google (outputs obtained in March, 2019). Both are generic MT systems, i.e., general purpose systems trained with huge amounts of data from various subject areas and thus suitable for translating texts in all subject areas or domains. Google Translate, in particular, is the best-known MT service, which can be used either free of charge as a standalone tool (translate.google.com) or for a small fee via an API for translating large amounts of text or for using it within a CAT tool.

2.4 Evaluation of MT outputs

In order to evaluate the quality of each MT output, we used both automatic evaluation metrics (AEMs), i.e., BLEU, METEOR, WER and TER, as well as human evaluation metrics (HEMs), i.e., side-by-side ranking, adequacy and fluency rating and error annotation

2.4.1 AEMs

The AEMs used in this study were BLEU, METEOR, WER and TER. BLEU is a score for the comparison of a candidate translation with one or more reference translations by measuring the number of n-grams (Papineni et al., 2002). It is the standard metric used in the MT community for a quick rough estimation of the quality of the MT engine performance. However, BLEU is relatively unintuitive and relies on a large set of MT outputs and a large set of human reference translations to correlate with human judgments (Snover et al., 2006). METEOR (Lavie and Agarwal, 2007) is based on the weighted harmonic mean of unigram precision and recall. WER (Zechner and Waibel, 2000) and TER are based on the Levenshtein distance and calculates the number of edits required to make an MT output match the reference translation. Since the use of a single human-translated reference tends to introduce bias (Popovic et al., 2016), two reference translations by professional translators were used.

2.4.2 HEMs

The human evaluation included side-by-side ranking of the MT outputs, adequacy and fluency rating and error classification. This evaluation was carried out by the ten MA students who performed the PE task, following completion of the PE task.

2.4.2.1 Side-by-side ranking

The evaluators were asked to read the Greek translations of each English source segment carefully and rank them in order from best to worst. The MT outputs were presented to the evaluators in a different order.
2.4.2.2 Adequacy and fluency rating

Following the ranking task, the evaluators were asked to rate each segment from each MT system for adequacy and fluency (defined as the extent to which a target segment was correct in the target language and reflected the meaning of the source segment) using a five-point Likert scale. They were asked to rate adequacy in response to the question “Is the MEANING of the English sentence kept in the translation?” A five-point Likert scale was used, where 1 was “Not at all”, 2 “Barely”, 3 “Partly”, 4 “Mostly” and 5 “Fully”. Similarly, the translators were asked to rate fluency in response to the question “Considering only GRAMMAR and SPELLING, the translated sentence is:”. Again, a five-point Likert scale was used where 1 was “Very poor”, 2 “Poor”, 3 “Fair”, 4 “Good” and 5 “Excellent”.

2.4.2.3 Error annotation

Finally, the evaluators were also asked to carry out an error classification task following very specific guidelines. They were asked to use an error typology which combined the subset of the Dynamic Quality Framework (DQF) and Multidimensional Quality Metrics (MQM) harmonized for MT analysis, as suggested by Lommel and Melby (2018), with the MQM error typology widely used in previous studies, mainly due to the flexibility of the error types and their granularity (Klubička et al., 2017; 2018; Carl and Báez, 2019). In particular, the participants were asked to classify the errors in each segment in the two main error categories and their subcategories: adequacy (addition, omission, mistranslation, untranslated text, terminology error) and fluency (grammar error, punctuation error, style error, spelling error and typo).

2.5 PE effort

According to Krings (2001), there are three categories of post-editing effort: (i) temporal effort, which refers to the time taken to post-edit a sentence to a particular quality level, which “is undoubtedly the most important aspect of post-editing from an economic perspective”, but “only the obvious external form of post-editing effort” (Krings, 2001:54), (ii) technical effort, which refers to keystroke and mouse activities such as deletions, insertions, and text re-ordering and (iii) cognitive effort, which refers to the “type and extent of those cognitive processes that must be activated in order to remedy a given deficiency in a machine translation” (Krings, 2001:179). Interestingly, temporal, technical, and cognitive effort do not necessarily correlate, since some errors in the MT output may be easily identified, but may require many edits, while other errors may require just a few keystrokes to be corrected, but involve considerable cognitive effort (Krings, 2001; Koponen, 2012; Koponen, Salmi and Nikulin, 2019).

The aforementioned MT outputs were presented to the ten participants in a random order. The latter were then asked to fully post-edit them. More specifically, they were asked to carry out the PE tasks while the temporal effort (total task time), the technical effort (keystrokes: insertions and deletions) and the cognitive effort (number of fixations, mean fixation duration and total gaze time) were registered using a Tobii X2-60 remote eye-tracker and the Translog-II software (Carl, 2012). The effectiveness of using eye-tracking as an MT evaluation technique has been demonstrated in previous studies (Doherty et al., 2010; Stymne et al., 2012; Doherty and O’Brien, 2014; Guzmán et al., 2015). Although using eye-tracking involves humans, it also eliminates much of the subjectivity involved in human evaluation of MT quality as the processes measured by eye-tracking are largely unconscious (Doherty et al., 2010).

Prior to the execution of the tasks, a group meeting was organised during which the participants were informed about the nature of the experiments, the task requirements and the general and task-specific guidelines they had to follow. The participants were asked to carry out full PE of the MT output generated by the two MT systems, following the task-specific
guidelines, namely, to retain as much raw MT output as possible, to transfer the message accurately, to correct any omissions and/or additions (at the sentence, phrase or word level), mistranslations, morphological errors, misspellings and typos, erroneous terminology and inconsistent use of terms, as well as incorrect punctuation if it interfered with the intended message, but to avoid introducing stylistic changes. On the day of the actual experiment, a warm-up session preceded the actual execution of the PE tasks to familiarise each participant with the procedure; the data from the warm-up task were not included in the ensuing analysis and discussion. During the experiment, the ST was displayed in the Translog software in the top half of the screen and the MT output in the bottom half. The participants were asked to carry out the tasks at the same speed as they normally work in their everyday work as translators; therefore, no time constraint was imposed. In addition, they worked directly on the MT output.

2.6 Reception study

Finally, the extrinsic quality of the post-edited texts, i.e. acceptability by the end-users in terms of readability and comprehensibility (Gouadec, 2010; Suojanen et al., 2015), was evaluated on the basis of a reception study with average readers/end users. More specifically, the post-edited texts, 20 texts in total, were presented in a random order to 120 readers/end-users (see below). Each reader was asked to read and evaluate one text, and then answer a few questions about the comprehensibility and the flow of the text using a 5-point Likert scale. Each reader was also asked to express their view on whether the text could be improved or not and how and whether they would like to read the rest of each text or not. Each text was evaluated by six (6) readers.

Question 1: The comprehensibility of the text was:
1. Very easy; 2. Easy; 3. Fairly easy; 4. Not very easy; 5. Not at all easy

Question 2: The flow of the text was:
1. Very good; 2. Good; 3. Fairly good; 4. Not very good; 5. Not good at all

Question 3: Do you think the text could be improved?

Question 4: If yes, how?

As shown in Table 3, the readers participating in this study were both female and male, with the female readers outnumbering the male readers. They belonged to various age groups and the majority had either an undergraduate or a postgraduate degree. In the study, a distinction is made between two reading groups: the heavy-reading group, whose participants are reading professionals, such as Greek translators, editors and proof-readers, and the light-reading group, whose jobs did not involve intensive reading, such as medical professionals, engineers and lawyers (Choi, 2016: i). On the basis of this distinction, 46 of the readers belonged to the heavy-reading group while 74 belonged to the light-reading group.
Table 3. Readers’ gender, age distribution, education level and familiarisation with translation

3 Findings and discussion

3.1 Evaluation of MT outputs

3.1.1 AEMs

The AEMs indicated that the quality of the MT output of both texts was comparable and good enough, hence suitable for PE. For BLUE and METEOR, the higher the scores, the better the quality of the MT output. Generally speaking, a score below 0.15 meant that the engine was not performing optimally and PE was not recommended as it would require a lot of effort to edit the translation and reach a publishable quality, while a score of 0.50 or above was a good score and meant that PE could be used to achieve publishable translation quality (Lavie, 2011). On the other hand, the lower the WER and TER scores, the better the quality of the MT output. A high WER or TER score implied that a translation would require many changes during post-editing (Munkova et al., 2020).

<table>
<thead>
<tr>
<th>Gender</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age distribution</td>
<td>18-25</td>
<td>18</td>
</tr>
<tr>
<td>Education level</td>
<td>High School Leaving Certificate</td>
<td>8</td>
</tr>
<tr>
<td>Studies or working experience in Translation</td>
<td>Student or undergraduate in Translation</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 4. AEMs per system per text

3.1.2 HEMs

3.1.2.1 Side-by-side ranking

As shown in Figure 1, the NMT output was ranked first by 80% and second by 20% of the evaluators, compared to the SMT output, which was ranked first by 38% and second by
62% of the evaluators. To assess the agreement between the annotators we computed Fleiss' kappa coefficient (Fleiss, 1971). Inter-annotator agreement showed fair agreement among the annotators (κ = 0.40).

![Figure 1. Average ranking percentage per system](image)

**3.1.2.2 Adequacy and fluency rating**

As shown in Figure 2, the NMT output fared slightly better than the SMT output as regards adequacy, or conveyance of meaning, and much better as regards fluency, or grammaticality. Inter-annotator agreement shows fair agreement among the annotators for fluency (κ = 0.29) and slight agreement for adequacy (κ = 0.10).

![Figure 2. Weighted average rating for adequacy and fluency per system](image)

**3.1.2.3 Error annotation**

As regards the number of errors (see Figure 3), the analysis revealed that the NMT output contained the lowest number of errors overall. However, the types of errors revealed that the NMT output contained fewer fluency errors but slightly more adequacy errors. More specifically, the NMT output contained fewer grammatical errors, but more additions and omissions. Inter-annotator agreement shows fair agreement among the annotators (κ = 0.22).

![Figure 3. Total number of errors per system](image)
3.2 Measurement of PE effort (temporal, technical, cognitive)

Temporal effort

As regards the temporal effort, we measured the average time (in minutes) the participants needed to post-edit each MT output. As shown in Figure 4, the MT output generated by the NMT system required less time ($M = 11.58, SD = 3.23$) for full PE than the MT output generated by the SMT system ($M = 14.92, SD = 4.72$). According to a two-tailed two-sample t-test, that difference in average task time was not statistically significant $t(16) = 1.61, p = 0.06$.

![Figure 4. Temporal effort: Mean and standard deviation of task duration per system](image)

Technical effort

Similarly, the participants performed fewer keystrokes (both insertions and deletions) while post-editing the NMT output ($M = 370, SD = 120$) than the keystrokes performed when post-editing the SMT output ($M = 574, SD = 201$). When post-editing the SMT output the number of insertions was slightly higher ($M = 299, SD = 108$) than the number of deletions ($M = 275, SD = 96$), while when post-editing the NMT output the number of insertions ($M = 187, SD = 69$) and the number of deletions ($M = 184, SD = 59$) was almost the same in both cases. According to a two-tailed two-sample t-test, that difference in keystrokes was statistically significant $t(16) = 2.52, p = 0.01$.

![Figure 5. Technical effort: Mean and standard deviation of keystrokes per system](image)
**Cognitive effort**

As regards the cognitive effort, in this study we measured the average fixation count, the mean fixation duration (in milliseconds) as well as the average total gaze time (in minutes). That is the sum of all fixation durations, on both areas of the screen (ST in the top half of the screen and MT output in the bottom half) to compare the cognitive effort expended by the translators when post-editing each MT output. The analysis revealed that the fixation count was higher when post-editing the SMT ($M = 1752, SD = 492$) output than the NMT output ($M = 1423, SD = 277$). According to a two-tailed two-sample t-test, the difference in fixation count was not statistically significant $t(16) = 1.53, p = 0.07$. Apart from the higher average fixation count, the SMT output also triggered longer gaze time ($M = 9.46, SD = 2.47$) compared to the NMT output ($M = 7.64, SD = 1.55$), with this difference being statistically significant $t(16) = 1.72, p = 0.05$. Finally, mean fixation duration was slightly higher when post-editing the SMT ($M = 327.12, SD = 38.82$) than the NMT output ($M = 323.37, SD = 34.94$), although this difference was not statistically significant $t(16) = 0.34, p = 0.37$. Looking at the distribution of visual attention between the ST and TT areas we also noticed that, when post-editing both the SMT and the NMT output, the fixation count, the mean fixation duration, and the gaze time were higher on the TT areas than on the ST areas (see Figures 6-8). The fact that much of the activity involved in the PE task takes place in the TT area has already been confirmed by previous studies (Mesa-Lao, 2014; Carl et al., 2011). This might be related to the fact that a “translation suggestion is already presented for post-editing, so less inspiration from looking at the source is needed” (Elming et al., 2014:161).

![Figure 6. Cognitive effort: Mean and standard deviation of fixation count per system](image1)

![Figure 7. Cognitive effort: Mean and standard deviation of total gaze time per system](image2)

![Figure 8. Cognitive effort: Mean and standard deviation of mean fixation duration per system](image3)
3.3 Reception study

As already explained, a reception study was carried out to evaluate the acceptability of the post-edited texts by the readers in terms of comprehensibility and readability. It should be mentioned that to assess the agreement between the readers we computed Fleiss’ kappa coefficient (Fleiss, 1971) and we found fair agreement among them in all cases.

As regards text comprehensibility, the analysis revealed that more than half of the readers, 57%, found the post-edited texts to be either “easy” or “fairly easy”, 11% found them “very easy”, while 32% found them either “not very easy” or “not at all easy”. Comparing the comprehensibility of the post-edited NMT texts and SMT texts, it is worth noting that the latter fared better than the former. 37% of readers found the post-edited SMT texts “fairly easy” as opposed to 25% for the post-edited NMT texts, and 18% of readers found the post-edited NMT texts “not at all easy” as opposed to 10% for the post-edited SMT texts.

Examining the percentages for heavy and light readers regarding the comprehensibility of each text and the weighted averages, we noticed that the comprehensibility of both the post-edited SMT and NMT texts was fairly high for both categories of readers. Interestingly though, no heavy readers found the post-edited NMT text “very easy” compared to 19% of light readers. In addition, 30% of heavy readers thought that the post-edited NMT text was “not at all easy” compared to only 11% of light readers.

![Figure 9. Average percentage of readers' view on the comprehensibility of the post-edited texts](image)

As regards readability, or the flow of the post-edited texts, the post-edited SMT texts were rated as “good” or “very good” by 25% of the readers, “fairly good” by 32% of the readers and “not very good” or “not good at all” by 43% of the readers, while the post-edited NMT texts were rated as “good” or “very good” by 17% of the readers, “fairly good” by 28% of the readers and “not very good” or “not good at all” by 55% of the readers.

According to the weighted average, the flow of the post-edited SMT texts was “fairly good” for both heavy and light readers, while the flow of the post-edited NMT texts was “fairly good” for the light readers and “not very good” for the heavy readers. As regards the post-edited NMT texts, it can be seen that no heavy readers found the flow of the post-edited NMT texts to be “good” compared to 16% of the light readers, while 39% of the heavy readers found the flow of the texts “not good at all” compared to 19% of the light readers. As far as the post-edited SMT texts are concerned, however only 22% of the heavy readers found the flow of the post-edited SMT texts to be “fairly good” compared to 22% of the light readers, while 30% of the heavy readers found the flow of the post-edited SMT texts “not good at all” compared to 14% of the light readers.
Apart from the questions about the comprehensibility and the readability or flow of the post-edited texts, the readers were also asked whether the texts could be improved and how. For the majority (73%), both the post-edited NMT texts and the post-edited SMT texts could be improved. Indeed, 75% of the readers said that the post-edited NMT texts should be improved, and 70% of the post-edited SMT texts readers said the same. In this case, a moderate agreement (around 60%) was found between heavy and light readers. More specifically, the post-edited SMT texts could be improved for around 70% of heavy and light readers, while the post-edited
NMT texts could be improved for around 90% of heavy and 70% light readers. The main points of dissatisfaction, for all reader types and post-edited texts, involved long and complicated syntactic structures, extensive use of passive voice and spelling errors. Lack of cohesion, both syntactic and lexical, was also mentioned by 40% of light and heavy readers. They highlighted the lack of conjunctions or the prolific use of the conjunction “and” and the use of awkward or atypical collocations. Another interesting finding is that 20% of the readers of the post-edited NMT texts—all of them heavy readers—asked whether the text they read was the result of machine translation and 5% went as far as suggesting that it should be thrown away and written all over again from scratch.

Figure 13. Average percentage of readers' view on the flow of the post-edited SMT texts

Figure 14. Average percentage of readers' view on the flow of the post-edited NMT texts

Figure 15. Average percentage of readers' view on whether the text could be improved or not
4 Conclusions

In a nutshell, the study’s findings indicate that PE of SMT output was more demanding in terms of temporal, technical and cognitive effort and also required more edits than PE of NMT output, that was also in line with the high ranking of the NMT output by the evaluators. Yet, the reception study revealed that the comprehensibility of both the post-edited SMT and NMT texts was “fairly high” for both heavy and light readers, while the flow of the post-edited NMT texts was deemed to be less satisfactory than the flow of post-edited SMT texts by both heavy and light readers, and especially by the heavy readers. This is an interesting and rather unexpected finding given the increased fluency of the NMT output. This might suggest that the improved fluency of the NMT output can mislead post-editors who under-edit and thus produce a final post-edited text that is not 100 percent fit-for-purpose. This is only an hypothesis and will be tested in future work involving a refined qualitative analysis of the NMT output and its PE version. Another interesting aspect that arises from the study is that, according to the readers, neither the post-edited SMT nor the post-edited NMT texts fully satisfied the readers’ expectations or needs, as they identified areas that could be improved, especially at the level of syntax and textual cohesion.
Acknowledgements

We would like to thank the HUBIC Lab for providing the Tobii X2-60 remote eye-tracker for the purposes of this study.

References


Zechner, Klaus and Alex Waibel. 2000. Minimizing word error rate in textual summaries of spoken language. 1st Meeting of the North American Chapter of the Association for Computational Linguistics.
A Report on the TC43 Workshop: Drafting effective machine translation post-editing guidelines

Viveta Gene  
PhD Candidate, Ionian University  
(Corfu)  
f20gken@ionio.gr

Lucía Guerrero  
Machine Translation Specialist,  
CPSL  
lguerrero@cpsl.com

ABSTRACT

It is commonly agreed that post-editing guidelines are key for a successful outcome in machine translation post-editing (MTPE) tasks. In spite of this fact, the current lack of an international standard or recommendations on how to draft such guidelines means each company decides whether to draft them or not, and what type of information they should include. This risks creating confusion amongst the post-editors.

With that in mind, Viveta Gene and Lucía Guerrero, members of the MTPE Training Special Interest Group steering committee (a GALA initiative in which stakeholders from academia, clients, LSPs and post-editors collaborate towards drafting a common protocol on post-editing training), conducted a workshop at the AsLing TC43 conference on how to design post-editing guidelines. This report summarizes their presentation at the workshop, the polls that were launched and the results and subsequent discussion. It also proposes best practices and a first draft template for writing post-editing guidelines.

1 Introduction

According to academic research and the best practices described by several organisations (to be explored later), post-editing (PE) guidelines can be defined as the set of specific instructions that the requesters of a post-editing service, such as machine translation (MT) buyers, (language service providers (LSPs) or researchers, prepare for the post-editors so that they know exactly what is expected of them in terms of quality, tools, the areas to focus on and other important aspects that define how the post-editing task is to be carried out and determine the final output. The examples below illustrate to what extent post-editing guidelines are deemed relevant in any machine translation post-editing (MTPE) workflow:

“So specifying PE [post-editing] guidelines involves deciding on text quality acceptance which, in turn, depends on aspects such as client expectations, turnaround time or document life cycle, among others.” (Díaz and Rico, 2012)

“Every post-editing project needs to have specific guidelines for translators to comply with, since the guidelines may help clients and LSPs to set clear expectations, and save time and effort for translators.” (Hu and Cadwell, 2016)

Overall, we can conclude that the two main objectives of post-editing guidelines are to transform the customer’s expectations into clear specifications and save time and effort for the post-editors.

In this paper we will consider post-editing guidelines as part of a post-editing assignment which can happen in two different scenarios:
• Post-editing in an isolated environment involving an interface with source/MTPE output without any interaction with a CAT Tool — a research environment for example, and
• Post-editing in a multimodal interface involving CAT Tools, meaning that the post-editing guidelines are part of an assignment containing instructions for the use of translation memories, glossaries, style guides, client instructions, delivery instructions and more.

2 Why do we need post-editing guidelines?

Clear guidance on how to perform the MTPE task, indeed any translation task, is a very basic need especially where there are assumptions and grey areas. For example, in ISO 18587, the standard for the post-editing of machine translation output, it is stated that full post-editing describes the “process of post-editing to obtain a product comparable to a product obtained by human translation”. The word ‘comparable’ raises many questions: How do we define ‘comparable’? Does ‘comparable’ mean ‘equal’? And ‘comparable’ in terms of what? Style is also one of the most controversial aspects in translation specifications, and often the main subject of arbitration. What sort of guidelines would make clear what is and what is not a preferential or stylistic change?

To support the idea that post-editing guidelines should be part of any post-editing assignment, the findings during our collaboration work in GALA (Globalization and Localization Association) will be presented in the following section.

2.1 The MTPE Training GALA SIG

The MTPE Training GALA Special Interest Group (SIG) is a collaborative initiative aiming to develop best practices in the training and preparation of professionals who handle the post-editing of machine translated content. The goals of the group are the following:

• Share experience, common practices and standards in the field of post-editor training;
• Identify the training needs for post-editors based on the experience from all parties (academia, clients, LSPs and post-editors);
• Develop a common Post-Editing Training Protocol.

With Viveta Gene and Lucía Guerrero as moderators, the GALA MTPE Training SIG has monthly calls in which they present a topic, launch polls and invite all members to engage in a discussion. Sometimes the moderators also arrange interviews with industry and academia representatives who share their views about the current month’s topic. GALA provides all SIG members access to Basecamp, an online platform where they can continue their discussions in between the monthly calls, suggest new topics and upload documents such as presentations, minutes or papers.

So far, the SIG has examined the following topics:

• Skills and competencies of post-editors
• Current setting
• The service of MTPE
• Current gaps and “pain points”
• Matching profile and skills of the post-editor
• Training model for professional translators
At the last meeting (October 2021) the topic was: “Guidelines for setting a basic post-editing workflow”. It is to be noted that, in this case, ‘guidelines’ must not be taken as instructions for the post-editors, but understood in a managerial context, i.e., as the aspects that the requesters of the post-editing service should consider when preparing a post-editing assignment. The moderators invited participants to share their ideas about what these guidelines could be, and then they were put to the vote. The following list ranks them by votes received:

<table>
<thead>
<tr>
<th>Idea</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>How much effort is expected and what shouldn’t be corrected</td>
<td>6</td>
</tr>
<tr>
<td>Client expectations for MTPE and cost savings</td>
<td>3</td>
</tr>
<tr>
<td>Error typology list/Expected MT behaviour/Errors to be reported</td>
<td>3</td>
</tr>
<tr>
<td>Relationship of budget to quality. Guidelines to link budget to quality issues</td>
<td>2</td>
</tr>
<tr>
<td>Make sure you have the expertise in the field and are able to meet the expectations before accepting an MTPE task</td>
<td>2</td>
</tr>
<tr>
<td>Don’t start MTPE without having completed an onboarding/training session for a given customer/project</td>
<td>2</td>
</tr>
<tr>
<td>Focus on the scope and purpose of the MTPE task</td>
<td>1</td>
</tr>
<tr>
<td>Always review adequacy and fluency, especially adequacy</td>
<td>1</td>
</tr>
<tr>
<td>Defining how to monitor the post-editing jobs with a 2nd linguist review</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: Most voted ideas to be included in MTPE guidelines (GALA MTPE Training SIG)

The three most voted ideas deal with what should or should not be corrected according to the client’s expectations, which shows that guidelines for the post-editors is foremost in all stakeholders’ minds. In the following sections we will analyse the current gaps and propose ways to address them.

3 Current gaps

As already pointed out, individuals who are in charge of preparing post-editing assignments and/or would like to write instructions face several challenges:

- The lack of internationally agreed standards on how to draft post-editing guidelines, or the inconsistency of existing guidelines;
- The restricted availability of existing post-editing guidelines, limited exclusively to the organisations in charge of writing these guidelines, and thus the significant variations among them;
- The absence of a clear distinction between post-editing guidelines, which are supposed to be very specific to a post-editing task, and general assignment instructions, resulting in an overlap of instructions.

More specifically, existing academic research on how to draft MTPE guidelines is very scarce, and most of the papers can be classified under any of the following categories⁵:

- Language-dependent;
- Domain-based;

⁵ See the references section for a list of the papers found.
• Outdated (only apply to rule-based or statistical MT);
• Comparative studies;
• Exclusively focused on the two levels of post-editing (light and full).

This means that someone trying to write guidelines for the first time will not find the models they need in current research, because the existing studies do not fit all scenarios.

3.1 Light vs full post-editing

The above “gap” referring to light vs full post-editing is also a focus point in this workshop. This topic was examined in the light of Nunziatini and Marg (2020), who propose a better definition of a post-editing level which is between full and light (‘medium post-editing’). They present comparison tables showing that standards and industry practices generally agree on the existence of two post-editing levels (light and full), but they have different views on what these entail. They suggest the idea that there is room to offer flexible post-editing services between those two, depending on the aspects to be focused on (e.g. technical terms, brands/product names, etc.).

We believe that this research confirms the assumption that simply asking post-editors to carry out full or light post-editing is not enough, some of the reasons being the following:

• **Misunderstanding about the scope.** There is a widespread assumption that light post-editing is a level which should be applied when the quality of the MT engine is so good that only a few tweaks in the target text are needed. This conception is wrong, as the distinction between full and light post-editing is not based on the amount of editing effort, but on the purpose of the translation and the quality requirements. As defined in ISO 18587, “[light post-editing is] normally used when the final text is not intended for publication and is mainly needed for information gisting, i.e., for rendering the main idea or point of the text. In this level of post-editing, the output shall be comprehensible and accurate but need not be stylistically adequate.”

• **Style** is also a controversial topic, as shown by the lack of agreement between existing standards and guidelines. This is pointed out by Nunziatini and Marg (2020) with respect to full post-editing:
  o According to the TAUS guidelines, the style “may not be as good as that achieved by a native-speaker human translator”;  
  o Sharon O’Brien (O’Brien, 2010) recommends that stylistic and textuality problems are ignored;  
  o ISO 18587 recommends that client’s stylistic guidelines are followed, and highlights that the style should be appropriate for the text type;  
  o GALA\(^6\) points out that the style should be consistent and appropriate.

• Post-editors, in particular very experienced translators who usually deal with high quality requirements and have less experience with post-editing, **tend to engage in full post-editing** especially when light post-editing instructions are not clear. The ability to focus on the correction of specific error categories only while not fixing the rest, even if they are detected, certainly requires clear instructions just as much as training.

---

Finally, it makes sense to conclude that insofar as there are many different parameters that define a translation assignment, the same should apply to post-editing. It makes sense to be able to offer ‘medium’ levels of post-editing, and not just full and light, which leads us again to the need to develop clear guidelines covering the specific aspects to focus on at each of these levels.

4 Addressing the gaps

All these shortcomings and challenges, if not addressed properly, can result in a negative perception of the guidelines by the post-editors, who may consider them:

• Too over-elaborated, creating extra effort;
• Too dense, i.e., lacking examples or practical information;
• Too long;
• Based on too many assumptions which should be explained;
• Repetitive or, even worse, inconsistent with the rest of the usual assignment instructions.

Considering the existing gaps, the different and sometimes contradictory assumptions that exist regarding post-editing levels, and the fact that each task has its specific requirements in terms of quality, it may be concluded that it is extremely important to summarize all these quality expectations in a document that the post-editors can understand, and which effectively can guide them towards meeting the quality expectations, while at the same time saving time and effort.

However, it seems impossible to draft one single instruction document covering all kinds of post-editing jobs and scenarios, not even in the case of all post-editing tasks for a particular language combination or quality level. What is needed is a flexible tool or template that will guide the stakeholders through all the aspects that should be considered by the post-editors when carrying out their assignments.

This idea is not completely new and has been explored before. EDI-TA was a R&D project funded by the European Commission as part of the MultilingualWeb-LT (Language Technology on the Web) group and conducted by Linguaserve and Universidad Europea de Madrid in 2012. The aim was to define a methodology for MTPE and included a strategy for inserting metadata into the bilingual files for later use, as well as a dynamic post-editing tool containing all the aspects which must be considered when writing post-editing guidelines and inserting the above-mentioned metadata.
Even if EDI-TA seems one of the most comprehensive and flexible approaches to drafting post-editing guidelines, it may be considered as outdated because it was based upon the MT output produced by a rule-based system (Lucy Software), and only a few language pairs were involved in the project (EN-ES, ES-EN, ES-CA, ES-EU). In addition, the concepts of rules and patterns are obsolete in the context of neural MT, as these concepts describe the behaviour of rule-based and statistical MT: that generally speaking, produce consistent errors, whereas NMT (neural machine translation) does not (the choice of a term may be correct in one segment, and incorrect in the next one). It seemed, though, an excellent starting point for this workshop.

5 Objectives of the workshop

The objective of the workshop was twofold:

- Create awareness of the importance of sending accurate instructions to the post-editors so that they may carry out the post-editing task according to the customer’s expectations, focusing exclusively on the aspects that are required.

- Collect feedback to create a flexible and granular template for writing effective post-editing guidelines in the context of NMT including:
  - Content
  - Length
  - Format
6 Polls

The participants were presented three polls in Slido\(^7\): one as a warm-up, to find out how many people were either (potential) writers or (potential) users of post-editing guidelines, and then two different sets of questions for each (writers or users). The questions in the latter two sets are similar but have a different perspective. The results of the polls are presented in the following tables.

6.1 Warm-up poll

The following answers imply that most of the attendees were both writers and users of post-editing guidelines. The reason for this, as some participants pointed out later in the discussion, is that sometimes post-editors end up writing the guidelines for their customers, especially for those who are less familiar with MTPE workflows.

![Warm-up poll chart](chart.png)

6.2 Writers’ and users’ polls

The questions in the polls seek to find out the participants’ point of view regarding the main aspects required in a flexible tool for drafting post-editing guidelines. These topics were selected because they are either not defined elsewhere or controversial: frequency, difficulty (for writers) / utility (for users), content, extension and format.

As regards frequency, most writers and users write/receive post-editing guidelines only in specific cases. However, a significant 30% of writers always send post-editing guidelines for all jobs and an equally significant 30% of writers never send guidelines, which seems to correlate with the 30% of users who never receive post-editing guidelines.

---

\(^7\) Slido (slido.com) is an online platform for launching polls, casting votes and seeing the results in real time.
The second question was a bit different for both writers and users. Most of the writers consider that drafting post-editing guidelines is difficult because they do not know what information they should include — which proves how relevant it is to define the content. The good news is that most users think that post-editing guidelines are useful, and only 33% consider that they are not for various reasons, the main one because they are too superfluous and vague. Writers should take good note about this. Finally, there is a discouraging 8% of writers who think the task is difficult because the users do not read them.
Both writer and user responses are aligned as regards extension: the threshold between what is acceptable and unacceptable as length seems to be 2 pages. Surprisingly, 17% of writers manage to fit the post-editing guidelines on a single page, and that is exactly the same percentage of users who consider that more than 1 page would be too long, which raises the question as to how much information can be conveyed on a single page.

Q2-Users: Do you find them useful?

- Yes: 67%
- No (they are too superfluous and vague): 25%
- No (others): 8%
- 0% (not applicable)

Q3-Writers: Which is the average extension?

- > 3 pages: 8%
- 2-3 pages: 17%
- < 1 page: 17%
- 1-2 pages: 58%

Q3-Users: Starting from which extension you would consider them too long?

- > 3 pages: 17%
- > 1 page: 17%
- > 2 pages: 67%
The polls show agreement on the most relevant content as well. The sections relating to level of quality expected and purpose of the translation received most votes in both cases, which makes perfect sense because both are interlinked — the purpose defines the level of quality. The remaining sections (tips, examples, subject area, content type, type of MT system) obtained a similar percentage distribution for both writers and users but received far fewer votes. Note that almost 20% of both writers and users replied ‘Other’.

Finally, the format for post-editing guidelines receiving most votes, from both writers and users, was the document (text processor, PDF) followed by the kick-off meeting. During the discussion some participants agreed that these two should not be considered as an alternative.
but as complementary. The rest of the options (spreadsheets, recorded videos, other) did not obtain any votes from the writers or the users.

7 Conclusions

The original idea for this workshop included launching an online questionnaire prior to the AsLing conference, with more questions, open to anyone involved in MTPE workflows and not only to AsLing participants. Having the replies before the workshop would have allowed the moderators to prepare an interactive discussion with the audience and start drafting collaboratively a template with the most important aspects to be considered when writing post-editing guidelines.

Given the framework in which the workshop took place, it was not possible to publish the questionnaire beforehand, as originally intended, and so the moderators adapted it to a shorter poll, with fewer questions, replied to during the session – thus with much-reduced participation and less time for discussion. Nevertheless, this workshop can be considered as a good starting point for further research on the topic. All participants agreed on the need to produce meaningful post-editing guidelines and their replies and comments allowed the authors to draw some conclusions in the form of recommendations and a suggested list of the sections to be considered when writing post-editing guidelines, which are presented below. These should be taken as only preliminary, to spark interest in further research.

7.1 Recommendations

- Define use cases for writing post-editing guidelines
- Plan for a kick-off meeting after sending the guidelines to the post-editors
- Make sure that they only contain the necessary information

7.2 Sections

- Subject area
• Content type
• Purpose of the translation
• Level of quality expected (light/medium/full post-editing)
• Type of MT system
• Examples of errors
• Tips (how to address the errors)
• Others (TBD)

8 Future work

Some of the questions that were raised and remained unanswered during the discussion due to lack of time were the following:

• Which aspects do writers take into consideration when deciding whether to write post-editing guidelines?
• Why do 8% of the writers think that users do not read them? How do they know, what are the consequences and what could be the cause?
• What content appears in the shorter guidelines (1 page maximum)? Is this sufficient for the users?
• What other sections do writers and users consider should be part of the post-editing guidelines?

With this in mind, the authors have agreed to continue the current work and are exploring options to take it further in the near future.

References


Integrating post-editing with Dragon speech recognizer: a use case in an international organization

Jeevanthi Liyanapathirana  
Fac. de traduction et d’interprétation  
University of Geneva  
juliyanapathirana@gmail.com

Pierrette Bouillon  
Fac. de traduction et d’interprétation  
University of Geneva  
pierrette.bouillon@unige.ch

ABSTRACT
In international organizations, the growing demand for translations has increased the need for post-editing. In this paper, we will explore the possibilities of using speech in the translation process by conducting a pilot post-editing experiment with three professional translators in an international organization. Our experiment consisted of comparing three translation methods: dictating the translation with machine translation (MT) as an inspiration, i.e., respeaking translation (RES), post-editing MT suggestions by typing (PET), and post-editing MT suggestions using speech (SPE). The speech recognizer used for this experiment was Dragon. BLEU and HTER scores were used to compare the three methods. Our study shows that the translators made more edits using the RES method, whereas with SPE the resulting translations were closer to the reference, according to the BLEU score, and required fewer edits. The time taken to translate was the shortest with SPE, followed by PET and RES methods. To the best of our knowledge, this is the first quantitative study to be conducted using post-editing and dictation together, where both reference translations, as well as revised, post-edited MT translation outputs, are used to perform a detailed analysis of the different possible translation methods that can be used in international organizations.

1 Introduction
In today’s world, automatic speech recognition (ASR) software contributes to the ergonomics, productivity and quality of many situations in our daily lives (Joscelyne, 2018). It can be found at the end of customer-support hotlines, is built into operating systems and is provided as an alternative text-input method for smartphones (Liyanapathirana et al., 2019). At the same time, improvements in machine translation (MT), along with the increasing demand for translation support, have allowed post-editing (PE) to become the norm in the translation industry. Post-editing MT can make it possible to achieve considerable savings, in time and cost, for large volumes of translation. As both of these areas (ASR and MT) reach higher levels of performance, thanks to technical developments, one can only anticipate that an interplay between the two would have the potential to create new business solutions and workflows (Mesa-Lao, 2014b), including, but not limited to, the translation industry.

Translation services are useful for a variety of industries and organizations, and especially among international organizations. International organizations, in general, require documents to be translated into multiple official languages with high accuracy, within a limited time frame. While most of these organizations have integrated MT support, very few studies have focused on the attitude of translators towards using MT, PE, and speech recognition in their translation workflows in this specific context (Liyanapathirana et al., 2019).

In this study, we conducted a preliminary investigation into the use of dictation and PE with three translators in a large international organization that volunteered to participate. We aim to analyze the difference, in temporal and technical effort, between using speech and typing to post-edit MT in the translation workflow of a large international organization. We also analyze whether spoken PE performs better when used to dictate the translation or when it is used to correct segments of MT. To our knowledge, this is the first detailed analysis of the
The paper is structured as follows: section 2 discusses work related to this study, while section 3 presents the profiles of the participants, the tools, and outlines the experimental design. Section 4 describes the experiments we conducted, as well as the data and results obtained. This is then followed by the discussion and conclusion in Section 5.

2 Related Work

In the context of machine-aided human translation and human-aided machine translation, different scenarios have been investigated in which human translators are brought into the loop by interacting with a computer through a variety of input modalities to improve the efficiency and accuracy of the translation process (Liyanapathirana et al., 2019).

While speech technologies such as Automatic Speech Recognition (ASR) and speech synthesis can improve the quality and efficiency of translator output, research and surveys show that they are not well integrated into current translation environments, and that a very limited number of translators and linguists make use of speech technologies in their work environment (Aparicio et al., 2001; CLoL et al., 2017; ELIA et al., 2018).

In the 1990s and 2000s, computational linguistics research investigated how to integrate ASR and MT, in particular by building text-to-text translation systems, where partial segments were spoken to incrementally correct the existing translation output until a final, high-quality translation was generated (Vidal et al., 2006). Subsequently, more and more research has been conducted on integrating ASR into the translation workflow: the performance of translation students and professionals when using commercial ASR systems has been assessed (Dragsted et al., 2011; Zapata, 2012); professional translators’ needs and opinions on ASR, as well as the possibility of using ASR in mobile and multimodal environments, have also been explored (Ciobanu, 2014, 2016, and 2018 and Zapata, 2016a, b). In recent years, researchers have been closely observing the use of speech recognition technologies by professional linguists (Ciobanu 2014; Ciobanu 2016; Ciobanu and Secară 2019). The findings of these studies have been used to enrich translator training programmes, which now often integrate speech-to-text techniques in their curriculum (Secară and Ciobanu 2019).

In terms of the potential use of ASR for post-editing purposes, several studies have been carried out (García-Martínez et al., 2014; Mesa-Lao, 2014a,b; Torres-Hostench et al., 2017, and Zapata et al., 2017). PE with the aid of a speech recognition system has been found to be the fastest method for translating (Zapata et al., 2017). Voice input has also been shown to be more effective than the keyboard alone for PE (García-Martínez et al., 2014). Finally, it has been found that the majority of translators welcome the possibility of integrating voice as one of the input modes for performing PE tasks (Mesa-Lao, 2014a, b).

While this previous research provides a solid background for investigating how speech technologies can be incorporated into the translation industry, very little research has been done on the combined use of ASR and MT in international organizations. In this specific context, in the 1960s and 1970s professional translators used dictaphones to read out their translations to transcriptionists. A preliminary survey involving six international organizations was also recently conducted to get an initial idea of the use of CAT tools, MT tools and ASR, and to assess the openness of professional translators to integrating ASR and MT in the translation workflow (Liyanapathirana et al., 2019). The results reveal that translators have a greater tendency to use dictaphones to translate from scratch, rather than speech recognition techniques, and that translators find speaking (e.g., dictaphones or other devices providing speech support) to be less tiring. Furthermore, if the ASR and MT techniques are of high quality, many
translators are open to the concept of using speech-based PE as part of their translation workflow. These findings pave the way for promising quantitative research, which may help determine whether productivity gains can be derived from speech-based PE.

These findings are the motivation behind our study on using speech and PE in an international organization. It is built around a use case, and further explores and expands on the research of Liyanapathirana et al., (2019) by using quantitative analysis, rather than a survey based on qualitative aspects.

The present study has two aims:

1. To explore how multiple modalities (speech and MT) can be integrated together within an international organization.

2. To understand how different methods of translation affect the translation process and performance (automatic comparison with both reference translations and revised outputs).

3 Methodology

We conducted quantitative research on the use of speech and PE in the trade domain in an international organization. Three professional translators were asked to translate three different texts (average length of 180 words) from the trade domain using three different methods: post-editing the MT suggestions by typing (PET), dictating the translation with MT as an inspiration, i.e., respeaking translation (RES), and post-editing the MT suggestion using speech (SPE).

3.1 Participant Profile

This preliminary study involved a large international organization based in Geneva that specializes in translating documents in the trade domain.

Three professional translators specialized in the trade domain, ranging between the ages of 30 and 35, participated in this study, as well as a professional translator who provided revision support. All participants had around 7+ years of translation experience. Their language skills involved translating from English to French. All translators were adept in post-editing via typing (PET) and using speech recognition support via Dragon. All participants were familiar with standard computer-aided translation software (SDL Trados). In addition, all three translators claimed to use speech input methods in their day-to-day life (to dictate messages in a smartphone or to issue commands to Google Home, Amazon Alexa, etc.).

3.2 Tools

Trados Studio (SDL, 2018) was used as the translation platform for translation and PE purposes (Figure 1). The MT suggestions were provided by NMT engines trained using the WIPO Translate software (WIPO, 2021).
Figure 1. A sample of Trados Studio: source text appears on the left side of the pane and target translation appears on the right side of the pane. Text is segmented at the sentence level. MT support (WIPO, 2021) appears above both panes, providing multiple suggestions for a given source sentence.

The speech recognizer selected for this experiment was Dragon Professional Version 14.00 (Dragon, 2021), as it could be easily integrated into Studio in the international organization’s environment (Figure 2).

3.3 Experiment Design

Three professional translators were asked to translate three different texts (of an average length of 180 words) from the trade domain using the following methods:

1. post-editing the MT suggestions by typing (PET)
2. dictating the translation using MT as an inspiration (RES). The translators could read the machine translation suggestions, but they had to dictate their entire translation using Dragon Professional.
3. post-editing the MT suggestion using speech (SPE). The translators could select a preferred machine translation suggestion out of multiple translation suggestions and edit the necessary words or phrases using speech commands via Dragon Professional.

Both RES and SPE required inserting and editing text with Dragon Professional using speech commands. ASR performance in Dragon was optimized for each participant. They were provided with additional training sessions to ensure that they were comfortable using Dragon Professional. They were shown how to add domain-specific vocabulary, use built-in commands already found within Dragon (Figure 3), as well as train new commands to navigate Studio using speech, when necessary. Figure 3 shows a subset of the commands featured in Dragon. The French commands in English are, respectively: “Select it”, “Deselect it”, “Cancel the previous action”, “Open the correction box” “Choose correction suggestion 1, 2 or 3” and “Correct it”.

Figure 2. Dragon Professional interface
Figure 3. Sample of Dragon built-in commands.

For example, the system was trained to add speech commands that enabled participants to switch between text segments within Studio. Figure 4 shows an example of custom-trained commands created by one participant to navigate within Studio (“Apply the second NMT suggestion”, “Select first NMT suggestion”, “Confirm translation and move cursor to next segment”, “Open dictation box”).

Figure 4. One participant’s custom-trained user commands for Dragon

All three translators were provided with the exact same text to be translated from English to French and the three translation PE methods were used successively (PET, RES, SPE). The time taken to complete each text was also logged.

Once the experiments had been completed, BLEU (Papineni et al., 2002) and Human Translation Error Rate (HTER) (Snover et al., 2006) scores were used to compare the translation performances for each of the three methods. This comparison was conducted in two parts:

1) Experiment 1: For the first part of the evaluation, the PE translations were assessed against the reference translations.

2) Experiment 2: For the second part of the experiment, the PE translations were compared against a revised post-edited MT translation. The revised translation was provided by a professional translator skilled in revising MT outputs. The following section describes and analyses the data collected.
4 Data Collection and results

4.1 Experiment 1: Comparison of translation outputs of different PE translation methods against reference translations

Below is an example of the source text, NMT output and a translator’s PE output. The differences between the NMT output and the translator’s PE output are highlighted in bold.

<table>
<thead>
<tr>
<th>Source Text:</th>
</tr>
</thead>
<tbody>
<tr>
<td>The system’s overriding purpose is to help trade flow as freely as possible — so long as there are no undesirable side effects — because this is important for economic development and well-being.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NMT output:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Le but primordial du système est d'aider le commerce aussi librement que possible, dans la mesure où il n'y a pas d'effets secondaires indésirables-parce que cela est important pour le développement économique et le bien-être.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Translator output:</th>
</tr>
</thead>
<tbody>
<tr>
<td>L’objectif fondamental du système est de contribuer autant que possible à la fluidité des échanges commerciaux, dans la mesure où cela n'engendre pas d'effets secondaires indésirables, car c'est un point important pour le développement et le bien-être économiques.</td>
</tr>
</tbody>
</table>

Once PE translations were collected from translators for all three methods, the average number of words generated by each PE translation method was calculated. Table 1 shows that Speech-based Post-editing (SPE) has a slightly higher number of target words per source word.

<table>
<thead>
<tr>
<th>Method Used</th>
<th>Average number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-editing via typing (PET)</td>
<td>1.24</td>
</tr>
<tr>
<td>Speaking the entire translation (RES)</td>
<td>1.25</td>
</tr>
<tr>
<td>Speech-based post-editing (SPE)</td>
<td>1.26</td>
</tr>
</tbody>
</table>

Table 1. Average number of words generated by each PE translation method (English-French)

PE translation performance for each of the three methods were compared against the reference translation using BLEU and HTER scores. Table 2 shows the average BLEU scores, HTER scores and the average time taken for each of the PE translation methods.
Table 2. Average BLEU score, HTER score and time taken in minutes for each PE translation method (English to French), and the HTER standard deviation.

<table>
<thead>
<tr>
<th>Method Used</th>
<th>Average BLEU score</th>
<th>Average HTER score</th>
<th>Average time taken</th>
<th>Standard deviation (HTER)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-editing via typing (PET)</td>
<td>38.43</td>
<td>0.485</td>
<td>28 mins</td>
<td>0.029</td>
</tr>
<tr>
<td>Speaking the entire translation (RES)</td>
<td>26.96</td>
<td>0.578</td>
<td>35 mins</td>
<td>0.066</td>
</tr>
<tr>
<td>Speech-based post-editing (SPE)</td>
<td>49.71</td>
<td>0.376</td>
<td>20 mins</td>
<td>0.007</td>
</tr>
</tbody>
</table>

Our results suggest that on average, the SPE method results in a better BLEU score and fewer edits to reach the reference translation, followed by PET, then RES. In addition, the time taken to translate was on average lower with SPE, followed by PET, then RES. This provides promising insight into the use of PE in general and, in particular, post-editing using speech.

4.2 Experiment 2: Comparison of PE translation outputs obtained from different translation methods against revised translation outputs

For the second experiment, the post-edited NMT outputs were revised by a professional translator and the revised outputs were used as the reference translation. Translation performances for each of the three methods were compared against the revised NMT translation using Translation Error Rate (HTER) scores. In addition, the types of edits made were also analysed in depth to get an idea of the types of edits necessary for each PE translation method.

Below is an example of the source text, NMT translation output, a translator’s PE output and a revised output. The main differences between the translator PE output and the reviser output are highlighted in bold.
The average number of words generated by the translators with each PE translation method, as well as the number of words in the revised translation output, can be found in Table 3, which shows that the translations obtained via the RES method result in a lower word count in the revised PE translation output. While the findings of this preliminary study may not be conclusive, it can be assumed that RES allows more freedom, and can result in an unnecessary increase in the number of words used during PE translation.

<table>
<thead>
<tr>
<th>Method Used</th>
<th>Average number of words by translators</th>
<th>Average number of words by reviser</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-editing via typing (PET)</td>
<td>258.33</td>
<td>246</td>
</tr>
<tr>
<td>Speaking the entire translation (RES)</td>
<td>187.66</td>
<td>174</td>
</tr>
<tr>
<td>Speech-based post-editing (SPE)</td>
<td>202.33</td>
<td>206</td>
</tr>
</tbody>
</table>

Table 3. Average number of words for each PE translation method and number of words in revised PE translation output. RES leads to a reduction in word count when it is revised.

For the next step, translation performances for each of the three methods were compared against the revised PE translation using BLEU and Translation Error Rate (HTER) scores. Figure 5 shows the average HTER score for each PE translation method, against the revised PE
translation output. The SPE method results in the fewest edits to reach the revised PE translation, followed by PET. RES requires most edits, which means that the reviser had to make more changes to the text.

Figure 5. Average HTER score for different PE translation methods between translator PE output and revised PE translation. SPE method results in fewer edits to reach the revised translation, followed by PET and RES.

When comparing the HTER scores of the PE translation outputs and the revised PE translation outputs (Figure 5), the number and types of edits that were made with the different PE translation methods were also analyzed using the HTER metric. Figure 6 shows the average number of edits and calculates the HTER score between the PE translation outputs and the revised PE translation outputs as a percentage of the text length. As expected, the RES method had the highest average edit percentage (compared to average text length), whereas the SPE method showed the lowest average edit percentage, thus in line with the results shown in Figure 5.

Figure 6. Average percentage of overall edits for different PE translation methods between translator output and revised translation. RES method showed the highest average edit percentage compared to average text length, whereas SPE method showed the lowest average edit percentage when compared to average text length.
In addition, when calculating the HTER score between the PE translation outputs and the revised PE translation outputs, a breakdown of the types of edits was also plotted.

![Avg. % of edit types with different types of translation methods (by reviser translation on translator outputs)](image)

Figure 7. Percentage of different types of edits when calculating HTER between PE translation outputs and revised PE translation outputs.

Figure 7 shows that all methods involved a high number of substitutions. The SPE method required the lowest number of insertions and the highest number of deletions compared to the other two methods. In the future, we hope to investigate the potential reasons behind these observations.

4.3 Impact of machine translation quality on methods based on post-editing (PET, SPE)

In this preliminary study, two of the three translation methods (PET and SPE) involved post-editing: SPE and PET. Both of these translation methods make use of machine translation suggestions, which means that the quality of the latter can have an impact on the performance of each of the methods.

An analysis of the edits made in the revised PE translation outputs was conducted in order to evaluate how many of those edits concerned words that appeared in machine translation suggestions. To calculate this number, we counted the number of words that were used by the translators simply because they appeared in the machine translation suggestions, but which differed from the reviser output.
In the above example, the translator used the words “font obligation”, which were found in the MT suggestion but which were then revised and changed to “imposent”.

The analysis shows that 51% of edits revised in PET outputs and 58% of edits revised in SPE outputs were found in the original MT outputs, demonstrating that MT has an obvious impact on word choice when it comes to post-editing whether typing or dictating. It also shows that SPE has the highest trace of MT suggestions, as compared to PET.

5 Discussion and Conclusions

In this preliminary study, we conducted a quantitative analysis of three translation methods: dictating the translation with MT as an inspiration (RES), post-editing the MT suggestions by typing (PET), and editing the MT suggestion by dictating (SPE). We did a quantitative analysis on the performance of each translation method using BLEU and HTER metrics. The first evaluation was carried out by comparing translator outputs against reference translations. The second part of the evaluation involved comparing translator outputs against revised PE outputs.

In general, the findings from this preliminary study suggest that professional translators working in an international organization can benefit from the integration of ASR for post-editing purposes.

The main findings of this study are:

1. When evaluated against reference translations, post-editing MT using speech (SPE) produces a better BLEU score, with fewer edits made compared to the other two methods (PET, RES).

2. When evaluated against reference translations, dictating the translation out loud (RES) results in the worst BLEU and HTER scores, suggesting that the changes do not improve quality.

3. Speech-based methods reduce the time taken to translate, as compared to typing.
4. Post-editing MT using speech (SPE) performs better when evaluated against revised PE translation outputs, resulting in a better BLEU score, as fewer edits are made compared to the other two methods (PET, RES).

5. In terms of the percentage of edits made, all methods involved a high number of substitutions when evaluated against revised translation outputs. The SPE method required the lowest number of insertions and highest number of deletions as compared to the other two methods.

6. High-quality ASR and MT support can increase the quality of a translation. The productivity gains of all three methods rely on the quality of the ASR and MT solutions.

7. Even though RES performed worse than the other techniques, the translators preferred both methods that used speech rather than typing (SPE and RES), as it allows them “to think out loud and in longer sentences”.

As this is a preliminary study, it has several limitations. Ideally, a study would involve more participants and longer texts. Despite these limitations, this use case provides interesting perspectives for future research. Post-editing using speech (SPE) led to higher BLEU scores, fewer edits and a reduction in the time needed to translate. Its performance is followed by PET and RES, respectively. This demonstrates that if ASR and MT output are of high quality and the translators are adept at using software (CAT tools, MT suggestions and ASR toolkits such as Dragon), speech-based post-editing is a promising approach that can result in performance gains in the translation workflow.

These findings and observations thus provide an initial step for us to move towards a more elaborate study involving a larger number of participants and a detailed evaluation of the types of edits that are made using each type of translation method. This preliminary study on how speech technology can be integrated in large international organizations using different types of translation methods thus concludes with positive and promising results.

Acknowledgements
We would like to thank our anonymous participants from the international organization (who provided translations), Lise Volkart from the University of Geneva (who provided all revisions of texts) and Danielle Thien from the University of Geneva (who provided proof-reading support for this article) for their valuable time and involvement in this pilot experiment.

References


KUDO Interpreter Assist: Automated Real-time Support for Remote Interpretation

Claudio Fantinuoli  
Mainz University/KUDO  
claudio@kudoway.com

Giulia Marchesini  
KUDO  
giulia@kudoway.com

David Landan  
KUDO  
david@kudoway.com

Lukas Horak  
KUDO  
lukas@kudoway.com

Abstract

High-quality interpretation requires linguistic and factual preparation as well as the ability to retrieve information in real-time. This situation becomes particularly relevant in the context of remote simultaneous interpreting (RSI), where time-to-event may be short, posing new challenges to professional interpreters and their commitment to delivering high-quality services. To mitigate these challenges, we present Interpreter Assist, a computer-assisted interpreting tool specifically designed for the integration in RSI scenarios. Interpreter Assist comprises two main feature sets: an automatic glossary creation tool and a real-time suggestion system. In this paper, we describe the overall design of our tool, its integration into the typical RSI workflow, and the results achieved in benchmark tests both in terms of the quality and relevance of glossary creation as well as in the precision and recall of the real-time suggestion feature.

1 Introduction

In recent years, the interest in computer-assisted interpreting (CAI) tools, especially in the domain of simultaneous interpretation, has significantly increased. CAI tools are desktop programs and mobile applications specifically designed to assist interpreters in at least one of the several sub-processes of interpretation: for example knowledge acquisition and management, lexicographic memorization, terminology access, and so forth. The tools available to date differ significantly both in the set of functionalities offered and in their architectures. They can be as simple as terminology management spreadsheets, available on the user’s computer, or complex applications deployed in the cloud, integrating simple features to support the daily work of interpreters or advanced functions that aim to automate some of the activities performed by interpreters (e.g., Fantinuoli, 2018).

CAI tools may be particularly relevant to support interpreters in their effort to maintain or increase the quality of their service in a changing professional landscape. This is particularly crucial given the recent rise in remote simultaneous interpretation (RSI), and the need to streamline processes, compensate for shorter time-to-events, and meet the constantly high demands in terms of interpretation quality. The transition from an analogue to a digital workspace, i.e. from a physical booth with a hardware console to the immateriality of an artificial environment, has favoured the interest in the integration of CAI tools into RSI consoles, increasing the willingness of practitioners to adopt novel technologies that can help them meet these new challenges. Recent advances in language-related applications, such as machine translation, speech recognition, and language modelling, have created the technical conditions for the design and integration of novel support features inside RSI workflows.

In this paper we present KUDO Interpreter Assist, a CAI tool specifically designed for the integration in RSI scenarios. Interpreter Assist comprises two main features: an automatic glossary creation tool and a real-time suggestion system. The goal of Interpreter Assist is to
shorten the preparation time and increase the precision of the rendition in highly specialized events.

The remainder of this paper is organized as follows. Section 2 introduces the related work in the area of CAI tools and the empirical experiments conducted on them so far. Section 3 describes the general architecture of the tool. Section 4 introduces the evaluation framework, the results of our tests, and the limitations of our evaluation methodology. Finally, section 5 concludes the paper and presents the outlook.

2 Related work

Computer-assisted interpreting tools have been proposed by several researchers in the past 20 years or so (Stoll, 2009; Fantinuoli, 2012; Will, 2015; Rütten, 2017;). Much attention has been devoted to the activities which are specific to the pre-event phase, i.e. the time prior to a conference that interpreters spend on linguistic and extra-linguistic preparation. In this phase, several automated or semi-automated approaches have been proposed, for example the collection of relevant documentation (Fantinuoli, 2017a), the extraction of specialized terminology (Fantinuoli 2006) and the exploration of corpora to investigate a new topic (Fantinuoli, 2017b; Xu, 2018).

A specific area of research has studied CAI tools as a means for retrieving specialized information while interpreting. Initially, this was done by manually searching for terms in a database and presenting results on an intuitive and distraction-free user interface. More recently, however, automatic speech recognition (ASR) has been proposed as a technology to automate the query system of CAI tools. This automation may solve the shortcomings of traditional tools, namely the excessive cognitive load required to use them, and extend the scope of information that can be retrieved by the machine (e.g., Prandi, 2017; Fantinuoli, 2017c; Hansen-Schirra, 2012). Not only terminology can be retrieved and presented to the interpreter in real time, but also other so-called problem triggers, such as numbers and proper names (Fantinuoli, 2017c; Rodriguez et al., 2021). Recently, researchers have started investigating the use of machine learning approaches for the automatic prediction of difficult parts of the speech (Vogler et al. 2019, Rodriguez et al. 2021).

Over the years, a handful of empirical studies have been carried out to test the feasibility of the human-machine interaction in the simultaneous modality. From a user perspective, they have mainly focused on the effectiveness of ASR-support during the interpretation of numbers (Desmet et al., 2018; Defrancq and Fantinuoli, 2020; Pisani and Fantinuoli, 2021), one of the problem triggers of simultaneous interpreting identified in literature (Braun and Clerici, 1996; Gile, 2009; Setton and Dawrant, 2016). Some have focused on the interpretation of terms (Van Cauwenberghe, 2020). In order to measure the impact on the quality of the rendition, these studies have used either mock-up systems (Desmet et al., 2018; Canali, 2019) or real-life tools (Defrancq and Fantinuoli, 2020; Pisani and Fantinuoli, 2021). Results seem to suggest that the use of automatic retrieval systems leads to increased precision in the rendition of the problem triggers under investigation. Little is yet known, however, about the influence of such tools on the overall performance of the interpreter.

CAI tools with integrated ASR developed to date comprise InterpretBank (Fantinuoli, 2017) and SmartTerp (Rodriguez et al., 2021).
3 Architecture

3.1 General workflow

Interpreter Assist comprises two main functionalities: automatic glossary creation and a real-time suggestion feature for terminology, numbers, and proper names. From the user perspective, the typical workflow comprises:

- Creation of a project with identification of the event topic
- Optional assessment of the automatically generated multilingual resources
- Sharing the resources with the team members
- RSI Interpretation with real-time suggestions

From the machine perspective, the process comprises the following steps:

- Creation of a mono- or multilingual domain-specific corpus
- Extraction of the monolingual terminology from the corpus
- Translation of the terminology into the target language(s)
- Fine-tuning of the baseline models with the generated resources
- Optional fine-tuning with client-generated information
- Real-time automatic access to the resources in the interpreter console

3.2 Glossary generation

The first step is corpus creation. The event-specific corpus is obtained using one of the three following methods:

- Using a set of documents in one or more languages that are relevant for a specific meeting provided by the user or client. These so-called preparation documents, generally provided by the event organizers, may comprise minutes of previous meetings, presentations, etc.

- Using a number of seed words describing a specific topic, for example “COVID-19”, “chronic respiratory disease” and “respiratory droplet”. Relevant documents are found by using these seeds with a general-purpose search engine, downloaded and added to the specialized corpus (cf. Baroni and Bernardini, 2004; Fantinuoli, 2006).

- Crawling a URL of interest, for example https://www.who.int/health-topics/coronavirus, used as a seed to crawl data from a given webpage. The text from the page is scraped and cleaned (removing headers, footers, meta) and saved. Links in the page are also extracted and followed to a specified depth either until there are no links left to crawl, or until the scraped text exceeds a default size. Preference is given to links with the same prefix and domain as the seed URL.

We generated term lists and scores using several approaches to terminology extraction (Ahmad et al., 1999; Bordea et al., 2013; Lefferts, 1995; Kozakov et al., 2004; Sclano and Velardi, 2007) on sample data similar to the data an interpreter might use in preparing for an event. After comparing the different results, we decided to retain the approach that offered the best compromise between quality of the term ranking and computational cost of running the algorithm. We chose to apply a modified version of the C-value algorithm (Frantzi et al., 2000).
The standard C-value algorithm ignores unigrams, so we introduced a constant that prevents the base-2 logarithm of term length (in words) from being zero for unigrams. For the part of speech (POS) patterns for candidate terms, we started with a standard pattern based on whether the head noun typically appeared in the initial or final position in a noun phrase, then modified the patterns after an initial analysis of candidate terms. The standard pattern for head-final noun phrases (e.g., English and German) is: 

\[(N \mid Adj)^* N \mid (N \mid Adj)^* N \! PP N)\],

and for head-initial noun phrases (e.g., French, Spanish, Italian) it is: 

\[(N (N \mid Adj)^*) \mid (N PP N (N \mid Adj)^*)\].

However, we modified both English and German after reviewing initial results so that the final POS pattern for English was: 

\[(N \mid Adj)^* N;\]

and for German was: 

\[(Adj^* N) \mid (N N)\].

We did not de-compound German nouns during processing.

Rather than converting the input text to lower case and lemmatizing, we saved considerations of case and morphology until all candidate terms had been collected. Thus, if a term appeared in the text in both singular and plural forms, we summed the counts and applied them to the singular form, while removing the plural from consideration; however, if only the plural form was encountered (e.g., big data), we kept the plural form rather than converting it to a singular (big datum).

Finally, we applied some heuristics and removed from consideration the following: terms containing certain stop words (typically comparative and superlative adjectives), unigrams common for the language, and terms that are left- or right-anchored subsets of a longer term. The top N terms and suggested translations (100 by default) were then presented to the interpreter, sorted in descending order of score. Higher scores indicated predicted usefulness of the term to the interpreter. The monolingual list of terms was then translated using in-house machine translation. The languages supported at the time of writing are Arabic, Chinese, English, French, German, Italian, Portuguese, Spanish, Japanese, and Russian.

The user then had the possibility of reviewing the generated resource, editing single entries or adding new terms. The resource can be shared with other team members.

### 3.3 Automatic suggestion

The automatic suggestion feature shows in real-time several units of interest in the online interpreter console, used during a remote interpretation session. Two kinds of information are shown to the interpreter:

- Specialized terms and their translations
- Entities such as numbers (with unit of measurement) and proper names (NER)

This feature is based on three main components concatenated in a cascading system. In the first step, automatic speech recognition (ASR) transcribes the speech in real-time. The ASR can be fine-tuned by using the data available for the project, namely the glossary itself, or information added by the client, such as a list of participants, product names and so forth. This operation helps to increase the precision of the tool, especially for out-of-domain words and proper names.

The transcription is sent to a language model (LM) for the identification of the units of interest: while the terminology is matched using the multilingual glossary generated (or a set of glossaries), the entities, i.e. the numerals and the proper names, are recognized directly by the language model in the ongoing transcription using NER. We decided to adopt different approaches for terms and entities following the recommendations of in-house interpreters. In a survey they had expressed the desire to be in full control of the terminology and the translations
displayed by the tool. While the chosen approach allows interpreters to trust blindly the suggestions offered by the tool, the shortcoming is that it does not offer help on terms that have not been curated by the user.

Once the results are retrieved in real-time, they are sent to the RSI platform interpreter console. To limit the abundance of information on the console, the user interface (UI) is designed to display suggestions in a non-intrusive way.

4 System evaluation

4.1 Dataset

To measure the performance of the tool we created two datasets, one to evaluate the terminology extracted and the other to assess the automated suggestion feature. For the purposes of this paper, both tests were carried out with English as a source language and Italian, French and Spanish as target languages.

To evaluate terminology extraction, we automatically generated three glossaries in three specialized domains: economics, industrial engineering, and medicine. All terminology lists were limited to 100 terms and were extracted from randomly selected .pdf files, which represent a typical format for preparatory documents in the context of interpretation.

For the automated suggestions feature, we created two datasets: the first comprised nine .wav audio files with general speeches, mostly covering the topics of politics, finance, and economics, but also including more generic topics and informal language (three audio files); the second dataset contained two longer .wav files comprising speeches with a higher density of specialized terminology. All files were public speeches and were representative of the typical speeches interpreted on the platform.

For every .wav file, two glossaries based on size were created: the first was a medium-sized topic-related glossary of 200 terms; the second was a big glossary comprising around 10,000 terms. They represented the two extremes in terms of glossary size and thus allowed us to test our terminology retrieval approach under different data conditions, assessing the impact of the underlying data structure on the tool’s usefulness, especially in terms of deterioration should too many unwanted results be retrieved (false positives).

4.2 Methodology

To test terminology extraction, we asked three professional interpreters to evaluate the English terminology list created by the tool. Evaluators had to assign each term to one of the following categories: specialized term, general term, or error. We define as “specialized” a term that is not necessarily understandable by people outside of the field, regardless of domain, while “general” indicates generic terms that most people can understand even when they pertain to a specific domain. “Error” marks incomplete or lexically invalid elements. We consider the “specialized” terms to be highly relevant for the creation of an event-glossary.

Using the same three glossaries, we then asked six interpreters (three for the language pair English>French and three for English>Italian) to evaluate the quality of the translation, marking each pair of extracted term and automatic translation as “acceptable” or “unacceptable” for the creation of a professional glossary to be used in an interpretation assignment.

For automated suggestions, we annotated each .wav file and prepared a list of all expected results in two categories: glossary terms and entities (numbers and proper names). We then ran
the tool with each speech and glossary (medium and long), generating an output made up of three elements: transcription of the speech, recognized entities, and terminology. We could then measure both precision and recall of the results, identifying issues caused by errors in the transcription and measuring the impact of false positives.

As far as named entities are concerned, results can fall into four categories: a) pass, when expected and correct; b) fail ASR, when not found because of an error in the transcription; c) fail REC, when correctly transcribed but not recognized by the LM; and d) false positives, when an unexpected term shows up among the results but is not present in the original speech.

As far as glossary terms are concerned, results can fall into six categories: a) pass; b) pass (different spelling), when the system correctly identifies variations in spelling of a same term; c) fail (different spelling), when the system fails to recognize a term because of a different spelling; d) fail (ASR); e) fail (term not matched), when the exact term is correctly transcribed but not recognized by the LM; f) fail (lemma not matched), when the expected term is different from the lemma in the glossary and is not recognized; and g) false positives.

4.3 Results

4.3.1 Evaluation of glossary generation

The evaluation of terminology extraction gave us an indication of how the list of automatically generated terms was perceived by interpreters. Across the three domains, the average categorization of terms indicated that the distribution between specialized and general terms was similar, with a slightly higher number of terms being considered specialized (Table 1). Terms perceived as errors represented around 1.5% of the total. This provided us with a generally positive evaluation of the tool, with very few errors and a tendency to select specialized terms for glossaries.

<table>
<thead>
<tr>
<th>Total Agreement</th>
<th>Partial Agreement</th>
<th>Disagreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>58.67%</td>
<td>41.33%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

*Table 4: agreements among interpreters*

An important value to put the results of terminology extraction into better perspective is the agreement among interpreters. This showed that total agreement was found in less than 60% of the cases, with partial agreement representing the remaining results. By “partial agreement” we refer to cases in which two out of three interpreters agreed on classifying a term as specialized, general, or error. Total disagreement, which would see evaluators marking a term under three different categories, was never found in our 300 examples. Total agreement seemed to be higher for specialized terms while partial agreement for general terms. This indicates a high variability in the way final users evaluate the level of specialization of single terms, probably depending on personal experience, past subject knowledge, etc. While our limited data only allows us to make an hypothesis, further experiments are needed to shed light on this.

<table>
<thead>
<tr>
<th>Correct Translations</th>
<th>EN&gt;FR</th>
<th>EN&gt;IT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Translations</td>
<td>91.2%</td>
<td>89.4%</td>
</tr>
<tr>
<td>Incorrect Translations</td>
<td>8.8%</td>
<td>10.6%</td>
</tr>
</tbody>
</table>

*Table 5: quality of term translations*

Translation quality was especially well received by interpreters. Our first attempt at evaluating translations from English to French and to Italian saw a very high percentage of terms considered correct for a professional interpreter glossary (Table 3).
4.3.2 Evaluation of automated suggestions

For the automated suggestions we used the data from the two tests to calculate precision and recall values for both entity recognition and glossary terms. Table 4 reports the average value among the nine .wav files containing the more general speeches. Both precision and recall were differentiated by glossary length (M=200, L=10,000). For term recognition a bigger glossary implies a higher number of false positives and, as a consequence, decrease in precision (from 98.99% to 88.82%). Recall values are relatively low on this general dataset. Typical errors in glossary term recognition were mostly related to a) 5-gram and short 1-gram terms (e.g., “right”, “fee”) and b) verbs, when their form did not correspond to the lemma (e.g., “addressed” for “address”, “studying” for “study”).

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Named Entities</td>
<td>89.83%</td>
<td>90.61%</td>
<td>89.89%</td>
</tr>
<tr>
<td>Glossary Terms (M)</td>
<td>98.99%</td>
<td>77.53%</td>
<td>86.58%</td>
</tr>
<tr>
<td>Glossary Terms (L)</td>
<td>88.82%</td>
<td>77.53%</td>
<td>82.49%</td>
</tr>
</tbody>
</table>

Table 4: general corpus without fine-tuning

Our second step addressed ASR fine-tuning as a means to improve the quality of the transcription and consequently of the retrieved results. We fine-tuned the models with a list of words – terms and proper names – simulating the availability of such information in the event organization pipeline. The results in Table 5 are promising. While not fixing the issues in their entirety, ASR fine-tuning allowed us to gain some percentage points in all categories, with a peak of 4.92% increase in the recall value of glossary terms.

The values presented above are average: the tool is generally able to maintain a high percentage of precision and recall in most of the test scenarios, with its highest results being: 100% (precision, recall, and F1) for named entity recognition; F1=97.78% for glossary (M) term recognition (P=100%, R=93.33%); and F1=96.55% for glossary (L) term recognition (P=100%, R=93.33%). The F1 value never dropped below 84% for named entities or below 81% for glossary terms, with the notable exception of one text: this particular sample, a generic speech about social issues, scored 76.19% (M) and 68.90% (L), with a higher number of errors and false positives impacting results. These preliminary results show that, while the tool tends to remain stable regardless of text or domain, end results can fluctuate depending on the terms we want to recognize and on the glossary chosen as reference.

Another factor we considered in honing results, in this case exclusively related to glossary terms, is the specificity of the lexicon. Since our hypothesis was that glossaries used by interpreters tend to be highly specialized, we performed a second test with two longer .wav audio files which presented a higher lexical density and more specific terminology. The results are presented in

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Named Entities</td>
<td>86.40%</td>
<td>92.68%</td>
<td>89.42%</td>
</tr>
<tr>
<td>Glossary Terms (M)</td>
<td>100.00%</td>
<td>96.30%</td>
<td>98.11%</td>
</tr>
<tr>
<td>Glossary Terms (L)</td>
<td>93.33%</td>
<td>96.30%</td>
<td>94.67%</td>
</tr>
</tbody>
</table>

Table 6: specialized corpus without fine-tuning
Table 6. Both precision and recall reached a better quality compared to the general language corpus, especially in terms of recall. On this dataset we once again applied ASR fine-tuning to improve the transcription quality. The results of this operation are presented in Table 7. They show a slight improvement from the baseline without fine-tuning. However, the effect of fine-tuning seems to be less prominent when compared with the more general corpus.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Named Entities</td>
<td>86.63%</td>
<td>94.19%</td>
<td>90.23%</td>
</tr>
<tr>
<td>Glossary Terms (M)</td>
<td>100.00%</td>
<td>96.30%</td>
<td>98.11%</td>
</tr>
<tr>
<td>Glossary Terms (L)</td>
<td>93.33%</td>
<td>96.30%</td>
<td>94.67%</td>
</tr>
</tbody>
</table>

Table 7: specialized corpus with fine-tuning

In the analysis of named entity recognition we paid specific attention to the category of numerals, which is particularly interesting in a real use case scenario. Our first test considered 95 numerals, and 93 (around 97%) were immediately recognized, and 2 ran into ASR errors that we were able to correct with ASR fine-tuning, thus reaching a pass rate of 100%. Our second test yielded slightly worse but still very good results, with a pass rate of 33/35 (around 94%). In this case 2 numerals encountered a recognition error, and because ASR fine-tuning did not aim at this particular entity type, no improvements could be achieved.

An important parameter related to the automated suggestions is system latency, i.e. how long it takes to process a term, recognize it, and display it on the screen. We conducted another specific test, taking into account the same nine .wav audio files we used for our first test, and selected three terms for each, both general and specific and of various lengths (x-grams). Measuring latency in seconds for each of these terms, we measured a response time that varied from a minimum of 1.1 seconds to a maximum of 2.3 seconds, with an average value of 1.6 seconds. This is within the typical ear-voice-span of interpreters and should be considered acceptable for the target use (Defrancq and Fantinuoli, 2020; Montecchio and Fantinuoli, 2021).

4.4 Limitations

The empirical results reported here should be considered in the light of some limitations.

Our evaluation was performed with a rather small and standardized benchmark test. For this reason, no generalizations on the performance of the tool can be made. Variables such as domain, speech features, accents, to name just a few, may have a huge impact on the performance of the tool and on its usefulness.

In the specific case of terminology extraction, agreement among interpreters deserves more attention, for example to determine acceptability thresholds. An important limitation to be considered is how these results, being evaluated by humans, are inherently subjective and depend on the evaluator’s perspective and background, as indicated by the low interrater agreement on term relevance. Input file type could be extended (e.g., with URLs), and a differentiation between domain-specific and non-domain-specific terms could be introduced.

Finally, testing automated suggestions only examined the situation at a specific moment in time: it could become iterative to consistently monitor changes in performance in terms of both precision and recall, and it would benefit from additional data.
5 Conclusions

In this paper we presented Interpreter Assist, a novel tool designed for simultaneous interpreters working in a remote setting. The tool offers support with the automatic creation of multilingual glossaries as well as with real-time suggestions of terms, numbers and proper names during the interpreting session. We built a dataset to benchmark the performance of the tool in real-life conditions, with typical speeches, topics and interpreters’ glossaries. Preliminary results are encouraging.

While the judgement of the relevance of the terms automatically extracted by the tool varied greatly among different evaluators, posing a considerable challenge to the tool’s ability to meet different user’s needs and expectations, the quality of the automatically translated terms was high. As far as the real-time suggestion feature is concerned, with a F1 value of around 98%, real-time suggestions performed well for specialized terminology and numerals, both with and without fine-tuning. However, recall values for general speeches still need to be improved. Performance for named entities increased considerably with fine-tuning. This supports the idea that an event-based pipeline to fine-tune the engines with event-related information is paramount to achieving high quality results. To successfully integrate this into the overall architecture of the tool, we aim to combine automatic fine-tuning of the models by means of the data obtained during the glossary preparation with a human-in-the-loop step which will allow users to add ad-hoc information related to the specific event.

References


Introduction

This paper reports on the panel discussion on *In-booth CAI tool support in conference interpreter training and education*, held at the 43rd Translating and the Computer Conference, the 8th organised by AsLing.

Computer-Assisted Interpreting (CAI) tools are software solutions designed to support interpreters in various sub-processes of interpreting with the aim of increasing productivity (e.g., by curbing preparation times) and output quality (cf. Fantinuoli, 2017; Prandi, 2017). The latest automatic speech recognition (ASR)- and artificial intelligence (AI)-powered versions of CAI tools for in-booth use provide interpreters with automatically generated visual aids for the interpretation of common “problem triggers” (Gile, 2009), i.e., items that are particularly prone to human error during interpreting, such as specialised terms, named entities and numbers.

This topic is of particular relevance because of the rapid and exponential growth in interpreting technologies in recent years and, predictably, in the years to come (Fantinuoli, 2018a). At the same time, the potential benefits of the use of CAI tools may help interpreters cope with changing contexts and market requirements. For instance, the growing demand for remote simultaneous interpreting (RSI), propelled by the Covid-19 pandemic, has made it necessary for interpreters to work at a distance. This limits the close collaboration inside the interpreting booth that has, so far, helped interpreters cope with challenges such as problem triggers in the source speech. On the other hand, interpreters may benefit from CAI tools in the context of a market characterised by lesser specialisation and shorter preparation time.

We can hence expect that the ability to leverage technological innovation to increase service quality and meet market demands represents a crucial “skill of the future” for conference interpreters.

Already before the pandemic it was considered that high-quality conference interpreter training should incorporate new technological solutions adapting to technological developments (Rodríguez Melchor, Horváth & Ferguson, 2020). However, in-booth CAI tool use is a novel approach to training and education.
area of professional practice, research, training and education. Given the novelty of the topic, scholarly discussions on whether CAI tool training is necessary, what it should entail and how to effectively train experienced as well as future interpreters appear limited.

The panel presented below provided a comprehensive discussion on the importance that different stakeholders from academia, the interpreting job market, professional associations and research ascribe to training on in-booth CAI tool use and its expected developments as well as key objectives for the future.

The paper starts with an introduction to in-booth CAI tools setting the background for the discussion. It then presents the panel’s aims, organisation and composition. Its key outcomes are presented by reporting panellists’ answers to the moderator’s questions. The paper then concludes with some final remarks.

**In-booth CAI tools**

The interpreting profession is undergoing a much-cited new “technological turn” (Fantinuoli, 2018b), characterised by the increasing penetration of information and communication technologies (ICT) into all phases of the interpreter’s workflow and the emergence of new technological applications for interpreting. Fantinuoli (2018a) proposes categorising such applications into two broad categories: “setting-oriented” and “process-oriented” interpreting technologies. Paraphrasing the author’s discussion, setting-oriented technologies (e.g., RSI platforms) influence the external conditions of interpreting. Process-oriented technologies (e.g., terminology-management systems and knowledge management tools) are designed to support the different sub-processes inherent to the interpreting activity, which may take place before, during, and after the assignment.

CAI tools are an example of a process-oriented technology aiming to support interpreters in various stages of their workflow and increase output quality and productivity. When CAI tools are used during the interpreting assignment to support the interpreter’s processing task, we speak of in-booth use. Note that ‘booth’ is used here in a figurative sense to include RSI assignments as well.

Several “generations” may be identified in the development of in-booth CAI tools (cf. EABM, 2021). Previously, in-booth CAI tool use consisted in the manual query of these software solutions to access terminology, in the same way as with a traditional digital glossary in Word or Excel (first generation) or with some further facilitation through advanced search algorithms (second generation). The recent integration of ASR and AI technology into CAI tools represented a major breakthrough. This novel technology, first introduced into InterpretBank ASR and more recently into SmarTerp and Kudo’s Interpreter Assist, paved the way for a third generation of CAI tools. Through a “cascaded system” of independent ASR and AI modules working in sequence (cf. Chuang et al., 2021), automatic transcription of the source speech is generated, ‘difficult items’ are recognised and ‘translated’ into a graphic representation displayed on the interpreter’s laptop screen in almost real-time. The final outcome is the display of problem triggers in a graphic format that is intended to facilitate the integration of the visual input into simultaneous interpreting (SI) (cf. Fantinuoli, 2017).

This third generation is hence characterised by the automation of the (previously manual) tool query for terminology. Another major difference from the previous generations is that third-generation CAI tools provide interpreters with support in the rendition of further common “problem triggers”, such as named entities and numbers. Within the framework of the SmarTerp innovation activity, a new “end-to-end system” characterised by an integrated speech and natural language processing model is currently being researched for its introduction into CAI tools, paving the way for a fourth generation (Gaido et al., 2021). This new system aims to
overcome some major limitations of third-generation CAI tools, including higher latency (i.e. the time-lapse between the utterance of the problem trigger and display of the visual aid) as well as error propagation (e.g. from the ASR to the AI module).

Looking at the evolution of CAI tools over the past two decades, we may consider the growth of these solutions to be exponential, i.e. “characterised by the doubling of its values at regular intervals” (Fantinuoli, 2019). Their number and impact may hence increase over time at an increasingly rapid pace.

The panel

Our panel focused on in-booth CAI tool support from the perspective of training and education. Its aim was to foster dialogue on the needs and challenges in this new area of educational practice and research.

The idea of organising the panel was first developed amongst the authors of the present paper, who all have an interest in the development of interpreting technologies and related research. The first author was appointed as the panel organiser. Given the limited expertise on the matter, the other two were selected among the panellists and later contributed to drafting the present paper.

The session took place on Tuesday 16 November 2021 and lasted for 1½ hours. It was held online due to the enduring Covid-19 pandemic. It was chaired by Susana Rodriguez, who introduced the panel topic, directed questions to the panellists, moderated the Q&A session at the end and closed the panel with final remarks.

To achieve its aims, the panel brought together high-level experts from interpreting training universities, the interpreting job market, professional associations and research on in-booth CAI tool use. The panellists were:

1. Dr Ildikó Horváth (Associate Professor and Head of the Department of Translation and Interpreting at ELTE University in Budapest, active freelance conference interpreter, President of the European Master’s in Conference Interpreting (EMCI) Consortium represented the views of European universities. The EMCI is a consortium of 16 European Universities founded by the Directorate General for Interpretation (DG SCIC of the European Commission and further developed in collaboration with the Directorate-General for Logistics and Interpretation for Conferences (DG LINC) of the European Parliament.

2. Javier Hernández Saseta (conference interpreter, Head of Multilingualism and Interpreter Training Support Unit at the Directorate-General for Interpretation of the European Commission) shared the views of a major employer of interpreters and their requirements in terms of the technological skills that interpreters will need in the future. SCIC interpreters provide their services at the Commission, the European Council, and other institutions. It is one of the three services that provide interpretation for the EU institutions, alongside DG LINC of the European Parliament and the Interpretation Directorate of the Court of Justice of the European Union.

3. Dr Alicja Okoniewska (experienced conference interpreter, trainer, researcher, member of the Research Committee of the International Association of Conference Interpreters (AIIC)) shared AIIC’s perspective on integrating CAI tools in interpreter training.

4. Dr Bianca Prandi (conference interpreter, trainer, PhD graduate from the University of Mainz/Germersheim, a leading researcher on process-oriented CAI tool research) shared her insights into empirical research on CAI tools.
5. Francesca Maria Frittella (conference interpreter, trainer, PhD candidate at Shanghai International Studies University working on the development of a research-based training solution on CAI tools through the integration of instructional design and evaluation into interpreting research) shared reflections on CAI tool training from an educational research perspective.

**Key outcomes of the panel discussion**

**The current state of in-booth CAI-tool training at EMCI universities**

The first question was directed to Dr Ildikó Horváth, President of the EMCI Consortium:

*What is the current state of training on in-booth CAI tools at EMCI universities? Is the use of CAI tools in the booth regarded as a future educational objective?*

Dr Horváth affirmed that CAI tools are considered a future educational need and explained that universities within the EMCI Consortium agreed to a common core curriculum and strive to promote quality standards in conference interpreter training. The core curriculum states that members have to constantly review changing needs, new developments and update their training programmes accordingly. According to the panellist, “In-booth CAI tool support is more and more thought about on the market. Such tools are already being developed, so it’s only natural that the EMCI has already started its work in this field as well”.

Dr Horváth explained that efforts to incorporate new ICTs into the training curriculum of EMCI universities typically starts with the identification of a need to proceed with the training of trainers (ToT). Examples mentioned were ToTs on RSI, introduced at the onset of the Covid-19 pandemic, followed by a ToT on CAI-tool enhanced assignment preparation. In the panellist’s view “it is essential for trainers to know these tools, their potential advantages and the obstacles that need to be overcome in their use, to guide training practice”.

**The current state of in-booth CAI-tool training at DG SCIC**

The second question was directed to Javier Saseta, Head of SCIC at the European Commission:

*What is the current state of training on in-booth CAI tools at SCIC?  
Is the use of CAI tools in the booth regarded as a future educational objective?*

In answering this question, Mr Saseta reminded us that DG SCIC is not a training institution but rather an employer of interpreters. As an employer, he added, DG SCIC is known as an excellence centre for the quality of interpretation service and the training of staff, which is particularly successful thanks to the availability of funding and consolidated training expertise. When it comes to collaboration with universities, Mr Saseta explained that SCIC’s role is to “orientate, point to market needs, set the requirements that interpreters need to fulfil to be ready to sit in our accreditation test.”

When it comes to CAI tool training, no solution has been developed yet. The DG SCIC turned to EMCI universities to ask what were their training advances concerning interpreting technologies. Future educational needs according to SCIC include the use of ASR technology, a user-friendly and comprehensive tool for remote interpreter training (for which SCIC launched a feasibility study) and an online selection tool.

**AIIC’s views on and support of in-booth CAI-tool training**

The third question was directed to Dr Alicja Okoniewska, a member of AIIC Research Committee:

*Does AIIC consider CAI tool training as a future educational objective?*
Is it promoting in any way the movement towards that objective?

Dr Okoniewska stressed the premise that the topic of CAI tool training is relevant not only for the new generations of interpreters but also for experienced professionals, who are represented by AIIC and are encouraged to embrace continuous professional development. In the panellist’s view, education and training should be research-based because “research tests and corrects the knowledge we have, identifies pitfalls and develops best practices.” In her view, the role of training is that of providing people with “a toolbox of skills” that they can employ once in the booth. Technology is considered as a new item to be added to the interpreter’s toolbox: “Last year AIIC acknowledged the pace of technological change and realised that new tools should be added to that toolbox, not just of new but also of very experienced interpreters”. In defining the development of CAI tool training as a future objective, however, the panellist added that a number of questions need to be addressed before that objective may be attained: “Where to start? How to start learning about CAI? Should we learn about the technology behind the tools? Should we learn about all tools on the market? Is it enough to learn to operate an interface to make the interpretation better?” She stressed that research can help respond to these questions, which is why AIIC set the use of AI, speech-to-text technology and CAI tools in the interpreting process as one of its key research priorities for establishing an annual grant. The other two key areas are RSI and the interpreter’s perception of new technologies in the face of the threat of a potential “de-humanisation” of new technologies, on the one hand, and to “re-humanise” technologies putting interpreters at the centre of technological development, on the other.

To support progress in these areas, AIIC established a Research Grant, worth 10,000 Swiss francs, in 2019 with the aim to bridge interpreting research and practice by commissioning studies of interest for interpreters and the whole profession.

State of the art of research

The fourth question was directed to Dr Bianca Prandi, who conducted her PhD research at the University of Mainz/Germersheim on the impact of CAI tools on the interpreter’s cognitive load:

What is the state of the art of empirical research on in-booth CAI tool use?

Dr Prandi highlighted that there are several strands of research on in-booth CAI tool use. A first strand focuses on the evaluation of the interpreting product, i.e., responding to the question: how does the use of CAI impact performance? Despite this strand of research being the most prolific, its focus has been rather narrow, i.e., centred on the rendition of isolated elements (such as numerals, terms, etc.). What was missing is a focus on other aspects of the delivery, such as the interpreter’s prosody, and the qualitative, communication-oriented analyses, which only recently started to emerge (Frittella, in review, in press). A second strand is aimed at the evaluation of a tool’s feasibility, i.e., the extent to which the CAI tool may be integrated into interpreting, as in a recent MA degree thesis on the impact of CAI tool latency on the interpreter’s delivery (Montecchio, 2021). A third strand focuses on the impact of CAI tools on the interpreting process, e.g., on the experienced cognitive load, as in the panellist’s PhD thesis (Prandi, in preparation), which is still a largely unexplored area of inquiry.

In general, the panellist highlighted a tendential shift from research on CAI tools with manual lookup to ASR-powered CAI tools and a growing interest in CAI-tool integrated RSI platforms.

Initial insights on possible educational needs

The fifth question was asked first to Dr Bianca Prandi, and then to Francesca M. Frittella:
What does research suggest about possible educational needs inherent to in-booth CAI tool use?

Frittella stressed the importance of this question: “Educational needs represent the starting point for the content of our training. Defining educational needs is, hence, the necessary starting point in developing a training solution.” However, both panellists highlighted that we currently are at a very early stage of research and, hence, the answer may only be tentative at present.

Frittella stressed that, in order to define educational needs, we should take into account “the great complexity hidden behind the apparent simplicity of AI-powered CAI tools. As trainers and researchers, we often realise that people tend to perceive CAI tools as systems that function independently of users. However, the fact that you do not need any manual input to operate the tool does not necessarily mean that interpreter-CAI interaction is ‘simple’ overall.”

The panellists argued that possible educational needs may emerge from negative patterns identified in previous studies. In Prandi’s words: “These patterns point to the fact that the situation may be more complex than what we tend to think and that the use of CAI tools may imply not only technical skills but also cognitive skills inherent to attention distribution, managing multiple sources of information, etc.”

Examples mentioned are the tendency to “over-rely” on the tool (cf. Defrancq & Fantinuoli, 2020), interpreting the individual problem trigger correctly but producing a sentence that does not make sense overall or omitting the whole following sentence, interpreters’ difficulty in distributing attention between the speaker and the CAI tool and balancing the interpreters’ agency with the necessary degree of trust in the tool.

Defining ‘CAI-tool training’

The sixth question was directed to Francesca M. Frittella:

What do we actually mean by ‘CAI tool training’? Is there a current definition?

She posited that “since the educational need, i.e. the basic content of our training, has not been defined yet, the very concept of ‘CAI tool training’ has not been defined either.”

She then provided a tentative definition of CAI tool training as: “Instructional measures aimed at ensuring that interpreters will use the tool effectively, i.e. to achieve consistently better performance than they would without.” She explained that this tentative definition rests on the assumptions that CAI tools have the potential to improve performance, human use of the tool influences the outcome of the interaction, and training should guide interpreters in enacting behaviour that is likely to ensure effective use of the tool.

According to the panellist, so far, we have only seen initial forms of CAI tool training advanced by tech-savvy professionals or introduced as a marginal part of research projects (e.g. Frittella, in review; Prandi, 2017, in preparation). Such initial forms of training have been limited to the provision of fundamental technical information on the tool’s technical specifications and user interface design, accompanied by the opportunity to test the tool in action. However, so far, no proposal has been advanced on how to guide interpreters in the effective use of CAI tools and support the skills and strategies inherent to this task.

Challenges to the integration into the EMCI curriculum

The seventh question was directed to Dr Horváth:

What are the challenges to the integration of in-booth CAI tool training into the EMCI curriculum?
The panellist highlighted the major challenges that lie in the dearth of research, trainers’ generally limited knowledge of available solutions and skills in using them, trainers’ attitudes to these tools, as in the perceived “dehumanisation” of interpreters, previously mentioned by Dr Okoniewska.

From a pedagogical standpoint, Dr Horváth highlighted the need to understand when to introduce CAI tools into the curriculum. She explained:

All training implies a progression scale following skills development. Different stages have different requirements in terms of educational objectives, assessment, feedback and the characteristics of the training materials provided (i.e. the source speeches). We are developing a complex set of skills in interpreter training and, with the introduction of CAI tools, probably a new set of skills need to be developed within this already complex set of skills. How and when should we be doing it?

She speculated that the right moment to introduce CAI tools into the curriculum is “not at the very beginning when the information-processing requirements are highest and the attention-sharing difficulty is at its peak level, but it’s not at the very end either” and called for research on this issue.

In summarising her answer, the panellist said that CAI tools are indeed considered a future educational need but research on these tools is needed to answer questions related to: “Who needs it? How do CAI tools influence cognitive load and information processing, anticipation, attention-sharing capacities? What is effective in-booth use? What is the concrete pedagogical methodology to use: when and how shall we train trainees?”

Skills of interest to the DG SCIC

The eighth question was directed to Hernández Saseta:

What kind of skills related to the use of CAI tools in the booth are of interest for an interpreter at the European Commission and why?

Saseta replied that SCIC has always been looking for “excellence” defined as: “Interpreters who are motivated, humble, able to accept feedback, can work in a team, are aware of ethical issues, ready to improve and accept new structures and situations.”

With regards to the new skills required in the future:

we will be looking for excellence plus, meaning that SCIC will need “interpreters who can be resilient, flexible, able to deal with uncertainty, familiarity with new technologies is fundamental, open and understanding of things, accepting new formats and able to incorporate a bigger, or perhaps a different cognitive load. We will be looking for such skills as well as an aptitude for comfortably navigating an increasingly multimedia and technological environment because our clients will need this as a service. This is the new situation which we are seeing every day.”

Research projects supported by AIIC for the development of required skills

The ninth question was directed to Dr Okoniewska:

Is AIIC currently supporting any research projects promoting the development of the skills that will be required of interpreters in the future?

Dr Okoniewska explained that the current selection for the 2021 AIIC Research Grant winner is still ongoing. Within the 2020 grant application, 14 projects were presented. The winner was “Inside the virtual booth” by Nicoletta Spinolo (Unibo) and Agnieszka Chmiel (Adam Mickiewicz University from Poznan, Poland). The authors proposed to explore how
different RSI environments influence interpreters’ emotions, perception, and performance with the aim to develop recommendations for platform developers.

**Future research needs**

The tenth and final question was directed to both Dr Prandi and Francesca M. Frittella:

*What type of research is needed to advance the state of CAI tool training?*

Dr Prandi stated that, in general, we need to broaden and deepen our understanding of human-computer interaction in the context of interpreting by expanding the scope of our research and focusing not only on the interpreting product but also on the process. We need to explore not just the micro-cognition of in-booth use but also the macro-cognition of, for instance, CAI-tool enhanced preparation, how this impacts in-booth use and the cognitive ergonomics of CAI tools. In her view, the latter is important because “we will increasingly use technology to optimise different ways of how we work, not just the use in the booth.” Finally, the panellist stressed the need to gather both quantitative and qualitative data and also to standardise our methods to obtain a sound body of knowledge guiding training practice.

Frittella addressed this question from the point of view of educational research, as her PhD project focuses on “the development of a training solution for the integration of CAI into SI from an Instructional Design perspective, so from the perspective of the science of how people learn and how to support learning processes through instruction.” She stressed that the development of a research-based intervention typically consists of (1) the identification of an educational need based on research, (2) the design of the intervention, (3) its development and (4) evaluation to improve the solution, on the one hand, and contribute to theoretical understanding, on the other.

From this point of view, research is needed for each stage of the development of a training solution, from empirical research on in-booth CAI tool use, to yield a precise definition of the educational need, to research on the intervention itself, to understand how to support interpreters in achieving an effective integration of CAI tools into SI.

In concluding, Frittella argued for a research-based approach to the development of a solution for in-booth CAI tool use:

Interpreter training has always drawn on the expertise of trainers, in terms both of the professional knowledge taught and the teaching methods used. But what happens if the field does not have that practical knowledge? This is the case for any new technological development. Because the solutions are new, we do not have a tradition of best practices, in terms of effective tool use and effective training strategies.

She highlighted that this problem is not limited to CAI tools and is likely to take many forms in the future:

We are now speaking of CAI tools but with the rapid pace of technological change, there will be new forms emerging at great speed. Training cannot wait for trainers to acquire sufficient practical knowledge to train others because this would mean persistently lagging behind new developments.

This is where, in her view, research comes into play to “anticipate a future need, develop a comprehensive, reliable understanding of the issue to generate a robust training solution.”

**Conclusion**

While in-booth CAI tool use is becoming an increasingly relevant phenomenon, the present panel has provided a valid comprehensive discussion on the role that these technologies...
are likely to play in interpreters’ continuous professional development and training at educational institutions in the years to come.

The discussions with high-level experts representing the EMCI Consortium, DG SCIC at the European Commission, the professional association AIIC, state-of-the-art empirical CAI research and educational research on in-booth CAI tool training allow us to summarise the panellists’ contributions on CAI tool training as follows.

Training on in-booth CAI tool use is considered a future priority across the realms of academia, the interpreting job market, professional associations and research. Universities belonging to the EMCI Consortium are considering integrating CAI tools into training, and some training-of-trainers seminars have already been directed to making new technological tools known to trainers. Regarding the EU institutions, technical dexterity as far as online platforms and CAI tools are concerned will represent a future prerequisite for interpreters. AIIC already considers skills related to AI and CAI tools, among other technologies, to be a necessary addition to the interpreter’s toolbox and therefore these are now focal points promoted through initiatives such as the annual AIIC Research Grant.

However, we are still at an initial stage of related research, which represents a challenge to the development of a training solution. Given the dearth of empirical evidence and the rather narrow and product-oriented focus of most of the previous research, the educational need, i.e. the skills that need to be developed for effective integration of CAI tools into the SI process, is yet to be defined. Because the educational need is as yet ill-defined, a definition of the very concept of ‘CAI-tool training’ and what it should entail has been missing, although a tentative definition is provided in the present paper. Because much complexity is hidden behind the apparent simplicity of CAI tools, more research is needed to broaden and deepen our understanding of this interaction. At the same time, research on interventions will be needed to foster the development of solid, reliable interventions and contribute to the understanding of how to effectively support interpreters’ development of the skills inherent to effective CAI tool use.

It is our hope that the results of these discussions will shed light on the challenges in integrating in-booth CAI tool support into conference interpreter training and help define goals for the short, medium and long terms, and will pave the way for further progress towards a pedagogically sound integration of in-booth CAI support into conference interpreter training.

References


ASR-CAI tool-supported SI of numbers:
Sit back, relax and enjoy interpreting?

Francesca Maria Frittella
Shanghai International Studies University
frittella@shisu.edu.cn

Abstract

Numbers represent the interpreting problem trigger “par excellence”. Recent studies suggest that automatic speech recognition (ASR)-powered computer-assisted interpreting (CAI) tools may support interpreters in the simultaneous interpretation (SI) of these complex and taxing elements. However, limited knowledge is as yet available about the extent to which ASR-CAI tools can support the SI of numbers and the factors that may determine effectiveness. This work contributes to deepening the understanding of this issue. Based on a theoretical framework and data from a case study, it will be argued that both machine factors (inherent to the tool) and human factors (subjective to the interpreter) may impact the delivery, with implications for both the design of ASR-CAI tools and training. The case study draws on a recent usability test of the ASR- and AI-powered CAI tool SmarTerp with 10 practising conference interpreters. The analysis reports recurrent and meaningful error patterns and discusses them in the light of the theoretical framework. The contribution may inform the work of others by presenting a novel approach to researching interpreting technology and developing recommendations to improve usability.

Introduction

Numbers may be regarded as the “interpreting problem trigger par excellence” (Frittella, 2019a, p. 79). In their review of studies on the simultaneous interpretation (SI) of numbers, Desmet and colleagues (2018) report alarming error rates hovering around 45–55% for students and 30–40% for professionals in experimental settings (cf. Desmet et al., 2018). Numbers also seem to be a source of stress for interpreters, as declared by 65% of the 826 respondents to a 1982 survey (Alessandrini, 1990). Some studies consider numbers to be even more problematic than other problem triggers (Lamberger-Felber, 2001) and even conjecture that common coping strategies do not apply to these elements (Pinochi, 2009).

Considering that numbers are so taxing, the growing interest in new forms of technological support for the SI of numbers seems only natural. One possible solution for the “trouble with numbers”, as Pellatt (2006) called it, seems to be offered by “third-generation” (EABM, 2021), in-booth computer-assisted interpreting (CAI) tools. CAI tools are software solutions “designed to support and facilitate some aspects of the interpreting task with the [primary] goal to increase [delivery] quality” (Fantinuoli, 2018, p. 4). Thanks to the recent integration of automatic speech recognition (ASR)- and artificial intelligence (AI)- into CAI tools (cf. Fantinuoli, 2017) — henceforth referred to as ASR-CAI tool support — these systems can now extract numbers from the source speech and display them on the interpreter’s laptop screen with minimal latency (cf. Fantinuoli, 2017).

Research on the potential benefits of using these technologies, from ASR mock-ups for numbers (Canali, 2019; Desmet et al., 2018) to CAI tool mockups (Frittella, in press-a) and live CAI tools (Defrancq & Fantinuoli, 2020; Pisani & Fantinuoli, 2021), is growing but still scarce. It is reported that error rates decrease by 22.5% (Defrancq & Fantinuoli, 2020, p. 19), 25% (Pisani & Fantinuoli, 2021, p. 195) and 30% (Desmet et al., 2018, p. 25) when ASR-CAI tool
support is provided compared to when it is not. This staggering reduction in error rates is even more remarkable considering that participants in the aforementioned studies (master degree students) did not receive systematic training prior to the study on how to use these technologies effectively in the booth.

Looking at these promising results, we may be tempted to conclude that the trouble with numbers is solved once and for all: from now on, to effectively cope with numbers, conference interpreters and students will simply need to turn on their ASR-CAI tool, sit back, relax, and enjoy error-free interpretation.

Unfortunately, the overall picture may be much more complex than that. Despite the promising reduction in error rates in the rendition of numerals, previous studies also report potentially negative effects of the use of ASR-powered CAI tools in the booth, such as participants’ perception of *distraction* and difficulty in *coordinating* the SI subprocesses (e.g. Pisani & Fantinuoli, 2021), as well as an apparent *overreliance* (Defrancq & Fantinuoli, 2020) on the tool. It has also been reported that the correct interpretation of a numeral may still be accompanied by errors in its context (Canali, 2019; Pisani & Fantinuoli, 2021). However, since most empirical explorations focused only on the rendition of the numeral (cf. Frittella, in press-a), few concrete examples are available that may help us understand the complexity of computer-assisted simultaneous interpretation (or CASI as in Prandi, 2022) of numbers, the problems that may arise in this task and their causes. A comprehensive understanding is essential both to improve existing tools and to instruct interpreters on how to leverage their potential.

Through the present paper, I aim to contribute to expanding the understanding CASI of numbers by exploring issues that may emerge in interpreters’ delivery. In discussing the potential causes of these issues, I will argue that both *machine factors* (i.e. those inherent to the tool’s technical specifications and interface design) and *human factors* (i.e. related to the interpreter’s skills) may impact the CASI of numbers, with implications for the design of ASR-CAI tools and training respectively.

To achieve this aim, I will draw on my recent usability test of the ASR- and AI-powered CAI tool *SmarTerp*8 (for more details cf. Frittella, in review) as a *case study* (cf. Swanborn, 2010). I will focus on how I used relevant literature to formulate principles for the design of the CAI tool, design the study, analyse and interpret data.

By its very nature, the case study presented is not intended to produce statistically valid results but rather to provide a different perspective on the research issue and unveil patterns of interest that may inform our discussions and provide directions for future research.

*Theoretical Framework*

*To what extent can a CAI tool support interpreters in the SI of numbers and how?* Considering that research on the CASI of numbers is still in its infancy, addressing this question is crucial to inform the design of CAI tools and related training. To effectively address this question empirically, an appropriate theoretical framework is needed to conceptualise the research issue, formulate hypotheses and interpret the findings.

At the current stage of research (e.g. considering that there is no model of the CASI of numbers and not enough evidence is available to develop one), the theoretical framework should draw on the extensive research that was conducted on the SI of numbers without ASR-

---

8 [www.smarter-interpreting.eu](http://www.smarter-interpreting.eu) [last accessed: 23/04/2022]
CAI tool support, and systematise the reasons why numbers are challenging elements to interpret. This is, however, a complex endeavour.

To provide a concise explanation, I will use the Processing Ladder Model (PLM) for the SI of numbers (Frittella, 2017, 2019a) as an organisational framework. I present this model which I developed in my previous work because to date it is the only model specific to the SI of numbers and it has been used in empirical research (Frittella, 2017, 2019a) and provided the foundation for the development of a training intervention on the SI of numbers (Frittella, 2019b).

The PLM was inspired by Chernov’s Probability Prediction Model (2004) in that it envisages the interpretation of numbers, like all interpreting, as a complex cognitive task involving several mental operations which may be conceptualised as *levels of processing*, represented in the figure below.

![Figure 1. Processing Ladder Model, adapted from Frittella (2017, 2019a)](image)

The model posits that processing on each level engages specific mental operations and that their failure may trigger specific errors. Processing failure may be caused by *mediating variables* inherent in the source speech increasing the processing requirements as well as the likelihood of error (for instance, the density of numerals in a given speech passage). However, failure may also be attributed to *idiosyncratic factors*, i.e. determined by the extent to which the individual interpreter possesses the *skills* (intended as both *automated skill components* and *interpreting strategies*, cf. Frittella, 2019b) required to successfully perform the mental operations inherent to each level of processing and limit the impact of possible mediating variables.

The idea that both mediating variables (objectively present in the source speech) and idiosyncratic factors (subjective to the individual interpreter) may influence the SI of numbers is consistent with Chernov’s conjecture that SI is influenced by the *redundancy* of a given speech unit which is both *objective* (arising from the characteristics of the unit at multiple levels) and *subjective* (determined by interpreters’ skills allowing them to recognise and exploit that objective redundancy). The higher the (objective and subjective) redundancy of the speech unit, the lower the processing requirements on the interpreter.

Because understanding the SI of numbers in all its complexity is crucial to move some steps forward in our conceptualisation of the CASI of numbers and identify the possible impact of technological support on this task, I will briefly explain the mechanisms inherent to each level of the PLM and revise supporting evidence. For a more detailed review, please see Frittella (2017, 2019a).

I. Numeral

The interpretation of numbers at its simplest involves translating numerals from the *source language* (SL) to the *target language* (TL). This task represents the first level of the PLM.
The task of ‘translating’ numerals may be best described as a form of bilingual numerical transcoding (Barrouillet et al., 2004). After the SL numeral is acoustically perceived, the information must be retained in memory while transcoding mental operations take place:

– **Decoding**, when a mental representation (e.g. acoustic or graphic) of the acoustically perceived SL numeral is generated.
– **Transcoding**, when the SL numeral is transformed into a graphic representation, e.g. as an Arabic numeral, either through visualisation or note-taking.
– **Recoding**, when the representation of the SL numeral, e.g. in graphic form, is turned into the TL numeral.

This level of processing in the SI of numbers is, arguably, the one that has been explored by the largest number of empirical studies. The major causes for errors at this level of processing may be summarised as follows:

– **Failure to acoustically perceive** the numeral: since, in most cases, numerals are difficult to reconstruct based on context (Mead, 2015, pp. 286–287), failure to pay sufficient attention to the numeral as it is pronounced by the speaker may lead to a loss of information (Pinochi, 2009, p. 40). This explains the fact that small numerals, characterised by shorter phonetic duration, are sometimes omitted (Defrancq & Fantinuoli, 2020, p. 24; Mazza, 2001) although they should be the easiest to process, as explained below.

– **Processing failure**, i.e. the failure in the mental processes of decoding, transcoding, and recoding. Processing failure may lead to several transcoding errors which essentially are of two types: **lexical**, i.e. when individual digits are misinterpreted (e.g. 15→50), and **syntactic**, i.e. when the relations between digits (e.g. 73→37) or their order of magnitude (e.g. 80,000,000→80,000,000,000) are not processed correctly. Note that several proposed classifications of errors in the SI of numbers (e.g. Braun & Clarici, 1996; Frittella, 2017) distinguish between lexical and syntactic errors.

– **Memory failure**: as the digits must be retained in working memory, which is quantitatively and temporally limited (e.g. Cowan, 2010), errors may occur as an effect of the working memory trace decay (Mazza, 2001). For this reason, in professional settings, professional interpreters either jot down figures or, more commonly, ask their boothmate to do so for them (Collard & Defrancq, 2019).

Empirical evidence suggests that transcoding numerals during SI is highly cognitively taxing for interpreters (Korpal & Stachowiak-Szymczak, 2018, 2019).

Mediating variables, which increase the difficulty of this task, seem to be a high number of digits and large magnitudes (e.g. numerals above one “thousand”) (e.g. Braun & Clarici, 1996; Mazza, 2001), as well as a divergent syntactic numeral structure between SL, graphic representation and TL (e.g. Pinochi, 2009).

Some authors conjectured that the mental operations of numerical decoding, transcoding and recoding may not be automatic for interpreters (Frittella, 2019a; Korpal & Stachowiak-Szymczak, 2018). In the PLM, the subjective degree of automatic transcoding operations, together with the interpreter’s ability to anticipate the upcoming numeral, listen attentively and use an effective strategy for memory retention, are considered to represent idiosyncratic factors influencing performance.

### II. Numerical Information Unit

During SI, a correctly recoded numeral must be linked to the components constituting a syntactic unit. For instance, consider the sentence:

Chinese export value decreased by 3 billion US dollars in 2019.
This task represents the second level of the PLM, in which all components of the numerical information unit (NIU) are processed and linked together. These are:

- **Numeral** (e.g. 3 billion)
- **Referent** — what the numeral refers to (e.g. export value)
- **Unit of measurement** — how the numeral is quantified (e.g. US dollars)
- **Relative value** — the qualitative variation described (e.g. decreased)
- **Time reference** (e.g. in 2019)
- **Geographical location** (e.g. in China)

Processing failure at this level may yield NIUs that are incomplete or incorrect, e.g. a numeral without a referent or attributed to the wrong location (Frittella, 2017, 2019a). In the case of multiple adjacent NIUs, the failure to correctly process all NIU components and their syntactic and semantic relations may lead to *misattribution errors*, where the interpreter incorrectly links a component of one NIU to another component of another NIU (ibid.).

The density of numerals within the numerical information unit is considered to be a mediating variable increasing processing requirements (cf. Mazza, 2001). More precisely, the PLM considers processing requirements to increase with the increase in (1) NIU components, (2) number of problem triggers in the NIU (e.g. if other components are highly technical specialised terms, acronyms, named entities or other numerals), (3) syntactic complexity of the NIU. Conceptually, the mediating variable may be identified with Chernov’s objective linguistic redundancy at the syntagmatic level.

The idiosyncratic component is modelled to arise from skills such as syntactic strategies reducing the complexity of the NIU and offloading the interpreter’s working memory.

**III. Text**

In processing the source speech, the interpreter must recognise numerals and NIUs as components of a cohesive whole and recreate the same semantic ties when producing a target speech. This task constitutes the third level of the PLM.

Processing failure at this level mainly manifests as numerals that are mutually contradictory and a target speech that lacks cohesion in the way numerals and NIUs are rendered (Frittella, 2017, 2019a).

Similarly to the NIU level, processing requirements at the text level are thought to be influenced by the degree of objective linguistic redundancy at the utterance level, in Chernov’s language, i.e. the number of numerals, NIUs and NIU components that are repeated at the speech level.

Analytical skills and specific strategies are seen as the idiosyncratic factor affecting the SI of numbers at this level. For instance, it was observed that, when interpreters fail to recognise the objective redundancy of the source speech and selectively omit repeated numerals and NIU components in highly dense passages, this recurrently leads to a series of errors called *snowball effect* (Frittella, 2019a).

**IV. Context**

The SI of numbers involves checking the processed information against relevant background knowledge to verify whether the interpreted numerical information is correct and plausible. This task represents the fourth level of the PLM.

If the interpreter is unsuccessful at this task, the delivery may present plausibility errors, i.e. blatantly implausible numbers such as “the world has 70.6 billion citizens” (Frittella, 2017, 2019a).
The unpredictability of the numerical information is conjectured to increase the risk for such errors.

The idiosyncratic component is represented by interpreters’ *encyclopaedic number knowledge*, i.e. their knowledge of pertinent numerical facts serving as benchmarks for the plausibility check (Frittella, 2017, 2019a).

### V. Function

The interpretation of numbers, like all interpreting, is not limited to transcoding units out of context but rather involves the transmission of a message situated within a specified communicative context. On the fifth and last level of the PLM, the interpreter identifies the communicative function of the numerical information in relation to the context and the goal of the speaker and recreates that same function in the target speech.

When this task fails, the interpreted numerical information is functionally inadequate even if its individual components were accurately rendered (Frittella, 2017, 2019a).

Processing on this level is conjectured to be influenced by the explicitness of the speaker’s intention, on the objective side, and by the interpreter’s analysis and inference skills, on the subjective side.

- **Conceptualisation of the CASI of Numbers and Design Principles**

  The theoretical framework above offers some conceptual tools to define the CASI of numbers, proffer expectations as to the extent to which an ASR-CAI tool may serve the interpreter and formulate design principles.

  The PLM considers the SI of numbers as a complex cognitive task involving several skills of varying complexity. Moving from this conceptualisation, we may posit that ASR-CAI tools (at their current stage of development at least) can support lower-level skills (i.e. transcoding numerals and the NIU) but not high-level skills (e.g. adopting strategies to overcome possible difficulties, producing an overall coherent speech and plausible numerical information).

  Consequently, design principles should be aimed at providing maximum support to interpreters in the transcoding of numerals and the NIU (levels I and II of the PLM).

  Starting from level I, the aim in presenting interpreters with the SL numeral in a graphic format is to minimise the risk of errors caused by a failure in acoustic perception, memory retention and numerical transcoding. To achieve this aim, we may expect that the numeral should be presented in a format that is as close as possible to the TL, hence preventing recoding errors. This leads me to formulate the first design principle (DP 1):

  **DP 1. Display the TL numeral in the graphic Arabic code according to the graphic conventions of the TL**

  Previous studies that presented participants with whole Arabic numerals (e.g. 3,000,000,000) or lexical elements of the numeral in the Arabic code and large orders of magnitude as a word in the source language (e.g. 3 billion) identified syntactic transcoding errors (e.g. ‘three million’) in participants’ deliveries (Canali, 2019; Pisani & Fantinuoli, 2021). To prevent similar errors, orders of magnitude above one ‘thousand’ should be displayed as in the TL graphic phonological code:

  **DP 2. Display orders of magnitude above one ‘thousand’ as a TL word**

  For PLM level II, to facilitate processing of the whole NIU, the CAI tool should display all NIU components. We may conjecture that the integrated presentation in a graphic unit may
help interpreters distinguish the boundaries between one NIU and the adjacent one and prevent misattribution errors.

**DP 3. Display all components of the NIU in the TL and as an integrated representation**

These design principles are aimed at providing interpreters with maximum support in the SI of numbers.

However, we may expect that patterns of errors can still emerge in delivery because, as explained earlier, an ASR-CAI tool displaying isolated numerals and some NIU components cannot support processing of the numerical information at the levels of its textual coherence, extralinguistic plausibility and functional equivalence to the original. We can hence expect that patterns of error at these levels may still emerge as an effect of idiosyncratic human factors, i.e. the individual interpreter’s access to the high-level skills that are responsible for processing the numerical information on level III and above of the PLM.

At the same time, to evaluate the extent to which an ASR-CAI tool can support the SI of numbers and evaluate the principles proposed for its design, the impact of mediating variables increasing processing requirements on the interpreter, as highlighted in the PLM, must be taken into account.

The considerations above have methodological implications (cf. Frittella, in press-a for a detailed discussion). Developing a comprehensive understanding of the CASI of numbers requires, first, to look beyond how the bare numeral was rendered and, second, to design the experiment to include speech passages of varying complexity, i.e. accounting for the mediating variables that were found to influence the SI of numbers, as explained earlier (cf. also ‘materials’ section below).

- **Method**

Building on the theoretical framework, I formulated design principles and proffered expectations on the extent to which an ASR-CAI tool may support the interpreter. In the analysis below, I will now present evidence from a case study related to these suppositions.

- **Background**

The data analysed in this paper is drawn from the usability test of the AI- and ASR-powered CAI tool SmarTerp conducted between January and August 2021. The individual live test sessions were conducted remotely due to the Covid-19 pandemic and moderated by the author of this paper. The reader may refer to Frittella (in review, in press) for more details.

This is chosen as a case study for this paper because the design of the materials is consistent with the PLM and the methodological considerations presented in Section 3. Furthermore, presenting delivery samples of practising conference interpreters selected with stringent criteria, rather than students, increases the robustness of the findings.

- **Materials**

Performance data was gathered through an SI test supported by a mock-up of SmarTerp. The test consisted in one speech being interpreted with the support of the CAI tool. The test speech was designed to include numerical tasks of varying complexity, i.e. speech passages containing different amounts of mediating variables increasing processing requirements on the interpreter, as defined by previous research on the SI of numbers (cf. section 2 ‘theoretical framework’).

SmarTerp’s support for numbers was designed following the design principles outlined earlier in the paper. However, it was not possible to follow DP 3 because, at the current stage
of its development, the AI engine could not recognise NIU components based on a semantic analysis but only extract elements recognised as problem triggers or named entities (for orders of magnitude). A screenshot of its interface may be seen below:

![SmarTerp's Interface](image)

**Figure 2. SmarTerp's Interface**

A mock-up was used both because SmarTerp was in an early development stage and the aim was to provide all participants with the same input, hence excluding the impact of uncontrolled variables. The tool’s peak performance and consistent latency of two seconds were simulated. This choice is consistent with the aim of the study to focus on design principles rather than the impact of other possible influencing variables, such as latency or tool failure.

- **Participants**

  The participants were 10 ITA (A) – ENG (B/C) conference interpreters recruited through an open call for participants and fulfilling the following selection criteria: 10+ years of professional experience, 30+ ITA-ENG SI workdays per year, 30+ RSI workdays over the past 12 months. Participant compensation was € 200,00. Before the test, all participants signed an informed consent and completed a self-paced online training module (delivered via the LMS Moodle) introducing them to the SmarTerp CAI tool to avoid first-time effects. The training module included basic information about the CAI tool’s UI and technical specifications as well as an exercise that was equivalent in structure to the test speech.

- **Data analysis**

  The analysis presented in this paper focuses on the most frequently recurring and significant patterns of error in the participants’ rendition of numerical tasks in the source speech and assesses them in the light of the theoretical framework as well as the considerations made in Section 3.

  Because the PLM conceptualises the SI of numbers as a multi-layered processing task, unlike previous studies (Defrancq & Fantinuoli, 2020; Desmet et al., 2018 inter alia), the evaluation of the delivery is not confined to the bare numeral but considers all levels of the PLM: the numeral, the NIU, the textual coherence of numerical information, its extralinguistic plausibility and functional equivalence to the original message.

  The classification of errors used, partly presented in the theoretical framework of this paper, was developed based on previous research on the SI of numbers (Frittella, 2017, 2019b) and discussed in depth in Frittella (in press-a).

- **Results and Discussion**

  Given the scarcity of insights regarding the impact of ASR-CAI on the SI of numbers beyond the rendition of the bare numeral and reflection on which factors may be conducive to errors, there is a risk of sketching an oversimplistic picture of the CASI of numbers.
The analysis of the data gathered during the SmarTerp test, used as a case study, reveals error patterns that may be interpreted as the interplay of both machine and human factors at all levels of processing of the input in the PLM.

In the discussion below, I will present and discuss some particularly relevant patterns. I will organise the discussion around evidence concerning the impact of the design principles formulated in Section 3. I will also present an interpretation focused on the impact of interpreters’ skills described in the PLM.

- **DP 1 and 2**

I formulated the design principles DP 1 (Display the TL numeral in the graphic Arabic code according to the graphic conventions of the TL) and DP 2 (Display orders of magnitude above one ‘thousand’ as a TL word) to provide interpreters with maximum support in numerical transcoding during SI, which typically involves the mental operation of decoding, transcoding and recoding (cf. section 2 ‘theoretical framework’).

Where these principles were not applied in previous studies (Canali, 2019; Pisani & Fantinuoli, 2021), syntactic transcoding errors occurred in participants’ deliveries, e.g. the visually presented ‘3 million’ was wrongly rendered as ‘3 billion’. These errors are predictable considering the hypothesis that numerical transcoding operations may not be automatic and that findings of previous research suggest that transcoding errors may occur at any processing stage, including recoding from graphic to TL numeral (Frittella, 2017, 2019a).

I hence conjectured that observing DP 1 and 2 in the design of SmarTerp would prevent the occurrence of such errors in interpreters’ deliveries.

The findings generally confirm this conjecture: no systematic occurrence\(^9\) of transcoding errors was detected in the dataset, with the single exception of the passage below which will be analysed in greater detail.

The delivery sample reported (which I translated as closely as possible from Italian) is representative of the error pattern identified in 6/10 deliveries:

<table>
<thead>
<tr>
<th>Task code</th>
<th>Source speech</th>
<th>Delivery (Molly)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NU</td>
<td>The continent currently has a gross domestic product of 3.42 trillion USD.</td>
<td>The continent has a GDP of 3.42 billion dollars.</td>
</tr>
</tbody>
</table>

\(^9\)The only exception were a couple of lexical transcoding errors in the delivery of one participant (e.g. 43 → forty-two), which I regarded as outliers.
Table 6. Error Pattern no 1 (6/10 Participants)

Looking at this error pattern from the point of view of possible interference of machine factors, we may explain the phenomenon as follows. The order of magnitude ‘trillion’ \((10^{12})\) was displayed as ‘bilioni’ in Italian on the CAI tool, which is the correct translation but rarely used. It also sounds like the English order of magnitude ‘billion’. An increasingly common alternative is ‘trilione’ \((10^{18})\), which is incorrect but used with increasing frequency as a loanword from English. Study participants misinterpreted this numeral, although it was correctly displayed by the CAI tool, because of the rarity of the order of magnitude ‘trillion’ and the ambiguity of its translation into Italian.

Interpreting the evidence from this point of view may lead to a revision of the design principles. I consider these findings to support DP 1 and DP 2, which I propose to consider valid until there is contrary evidence. However, the confusion caused by the display of ‘trillion’ points to the fact that rare orders of magnitude may be confusing to users, especially if several TL translations are possible and ambiguous. In similar cases, it may be recommended to (a) test several options with interpreters to find out which one is generally accepted as most appropriate within a given language combination, and (b) explicitly inform interpreters of the conversion system for orders of magnitude in the ASR-CAI tool.

At the same time, however, the error pattern may be interpreted as a consequence of the interference of idiosyncratic factors. First, the 6/10 interpreters who made that error did not know about all the possible translations of the order of magnitude ‘trillion’, as emerged from the interviews. Second, they did perform a plausibility check, leading to an error at level IV of the PLM of the extralinguistic context. It is, in fact, highly implausible that the whole African continent might have a GDP that is as low as the smallest economies in the world.

DP 3

The SI of numbers cannot be reduced to interpreting numerals out of context. Therefore, I formulated DP 3 (Display all components of the NIU in the TL and an integrated representation) to support interpreters in processing all components of the NIU.

Previous research reported that, when ASR-CAI tool support is provided for numbers, errors may still occur in the number’s context (Canali, 2019; Pisani & Fantinuoli, 2021). These findings are consistent with the errors identified in the SI of numbers without CAI tool support. I expected that DP 3 would help interpreters render the NIU completely and accurately, preventing errors such as open sentences (due to one or more missing NIU components) and inaccuracies in the number’s context.

However, this principle was not technically feasible and SmarTerp could only display components recognised as problem triggers – i.e., numbers, acronyms, named entities and, in
this case, specialised terms. Because ‘diamond production’ is not a specialised term but rather a general one, this referent could not be displayed.

In the table below, which reports the task ‘non-redundant number cluster’ (NCN), source speech elements that were displayed by the tool are highlighted in grey. The delivery sample is representative of the error pattern in 6/10 deliveries, which may be defined as a ‘misattribution error’ because the referent of the first NIU (‘nickel’) was wrongly attributed to the adjacent NIU too.

<table>
<thead>
<tr>
<th>Task code</th>
<th>Source speech</th>
<th>Delivery (Mermaid)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCN</td>
<td>Madagascar alone produced</td>
<td>Madagascar alone</td>
</tr>
<tr>
<td></td>
<td>approximately 58,000 metric tons</td>
<td>has produced 58,000 tons</td>
</tr>
<tr>
<td></td>
<td>of nickel in 2021</td>
<td>of nickel in 2021</td>
</tr>
<tr>
<td></td>
<td>Namibia’s diamond production</td>
<td>Namibia</td>
</tr>
<tr>
<td></td>
<td>amounted to 2.52 million carats</td>
<td>[has produced] 2.52 million</td>
</tr>
<tr>
<td></td>
<td>in 2018</td>
<td>of nickel</td>
</tr>
</tbody>
</table>

Table 7. Error Pattern no 2 (6/10 Participants)

From the point of view of machine factors, the error pattern may be traced back to the CAI tool not displaying the referent of the second NIU (‘diamond production’), while the referent of the first NIU (‘nickel’) was still displayed, as shown in the figure below:
This error pattern may hence be interpreted as evidence for the need to provide interpreters with all the components of the NIU, as formulated in DP 3. Although, as explained earlier, this was not possible from a technical standpoint, future dedicated studies may test the impact of the presentation of the NIU on the interpretation.

However, from the point of view of human factors, the error pattern may also be interpreted as the interpreters’ failure to be attentive to the tool and the source speech at the same time. Hence, it may represent an instance of ‘overreliance’ on the CAI tool (Defrancq & Fantinuoli, 2020 inter alia) as participants were using the CAI tool, rather than the source speech, as the primary source of information. Furthermore, the delivery sample above contains a sentence fragment since the unit of measurement of the second NIU (‘carats’) is missing. Due to this omission, the interpreted sentence is incomplete from a syntactic point of view. In the PLM, we may interpret this as the interpreter’s failure to activate analysis skills at the NIU level that would have allowed her to recognise all components and monitor the completeness of her delivery.

In the same way, the error pattern below, identified in 5/10 deliveries, shows an instance of a NIU component (the relative value ‘add’) that was not displayed by SmarTerp and was misinterpreted by study participants:

<table>
<thead>
<tr>
<th>Task code</th>
<th>Source speech</th>
<th>Delivery (Mermaid)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCR</td>
<td>By 2030, the African continent would add about 295 million new people aged 15–to–64.</td>
<td>By 2030, the African continent will have about 295 million people aged 15–to–64.</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>By 2030, Africa will hence be home to nearly 1 billion people aged 15 to 64.</td>
<td>And by 2030 there will hence be 1 billion people aged 15 to 64.</td>
</tr>
</tbody>
</table>

Table 8. Error Pattern no 3 (5/10 Participants)

However, the error pattern may also be interpreted as an internal inconsistency in the delivery, as the first and the second statements made by the interpreter are mutually exclusive, as well as the interpreter’s failure to check the first information for plausibility (it is highly...
implausible that the African working-age population will decrease in 2030 compared to today). Finally, this pattern may also be interpreted as a manifestation of the snowball effect that was identified in the SI of numbers without ASR-CAI tool support: in this highly dense speech passage, rather than omitting repeated NIU components to save time and dedicate more attention to analysing the source speech and its delivery, the repetition of redundant components caused an increase in processing requirements leading to severe errors.

• Conclusion

Technology is likely to play an increasingly important role for the profession in the years to come, especially regarding aspects of interpreting that impose a high emotional and cognitive burden on the interpreter and are particularly prone to human error. Seen as the “problem trigger par excellence” (Frittella, 2019a), numbers are receiving growing attention in the development of ASR-CAI tools and related research. However, research on the CASI of numbers is still in its infancy. Furthermore, the apparent simplicity of ASR-CAI tools threatens to produce an oversimplistic representation of the issue.

The present paper aimed to contribute to deepening understanding of the CASI of numbers to inform future studies. This aim was pursued through a review of relevant literature and the analysis of interpreters’ delivery with the AI- and ASR-enhanced CAI tool SmarTerp as a case study. Recurring and relevant error patterns were selected to discuss the possible influence of both machine factors, related to the CAI tool and, specifically, the design principles formulated, and idiosyncratic human factors determined by the individual interpreter’s readiness to activate skills and strategies allowing them to overcome challenges in the interpreting task.

The findings should be considered within the limitations of the case study method which, by its very nature, cannot produce statistically relevant results that may be generalized to a wider population. The study was also not intended to provide an exhaustive description of the research issue, but rather to propose a new approach and shed light on aspects that could advance the understanding and serve as a starting point for future research. Elements of innovation in the study include the ad-hoc designed test speech, constituted of numerical tasks of varying complexity, the in-depth qualitative analysis of error patterns and the focus on the identification of both influencing human factors and machine factors, leading to implications for CAI tool UI design and training.

In highlighting the complexity of the CASI of numbers, the study demonstrated the importance of framing empirical enquiry into a broader theoretical framework and using methods capable of capturing such complexity.

The discussion of a possible impact of design principles on the interpreter encourages progress in the study of CAI tool usability which, so far, has been dealt with rather marginally in empirical explorations.

Finally, the possible impact of human factors calls for further exploration of the role of skills and strategies in the use of CAI tools as a support for the SI of numbers. The idea that an interpreter may simply need to sit back and relax when an ASR-CAI tool is provided appears to be a fallacy that confuses the automation of the tool with the simplicity of its use. It is much more likely that the effective integration of the tool into the SI of numbers will require the activation of complex skills.
Funding

The data presented in this paper was collected within the SmarTerp innovation activity funded by EIT Digital. I did not receive any funding to author this paper or any other related scientific publications.

Acknowledgements

I owe much gratitude to SmarTerp, particularly Susana Rodríguez, for their support in the development of the materials used in this study.

References


Development of technological competences: remote simultaneous interpreting explored

Raquel Lázaro Gutiérrez
Universidad de Alcalá
raquel.lazaro@uah.es

Gabriel Cabrera Méndez
Universidad de Alcalá / Dualia
Teletraducciones
gabriel.cabrera@uah.es

Abstract
The technology applied to conference interpreting has always fallen behind progress. Only the booths, receivers, infra-red radiators and interpreting consoles improved with new versions, and a simple change from analog to digital technology took place over many years in the profession. However, the COVID-19 pandemic caused the efforts devoted to improve remote simultaneous and consecutive interpreting to skyrocket, and so videoconferencing is more and more present in our daily lives, and international events are using different platforms to connect people from all around the world. These platforms are slowly incorporating plugins that allow for the provision of interpretation, and improvements are made daily to ensure satisfactory results.

Professional remote interpreters complain about the technical conditions in terms of sound, connectivity and video that make their job much more difficult than before the appearance of SARS-CoV-2. For these reasons, we want to propose a contribution to the improvement of interpreters’ occupational health through the design of a training module that includes a series of robust and specific technical recommendations, supported by active professional interpreters and telecommunication engineers, in terms of sound, hardware, connectivity and the implications on computer performance or health, among other issues that must be included while training current and new professionals.

1 Introduction
The first reference to remote interpreting is said to be that of Paneth (1957), who, referring to remote simultaneous conference interpreting, characterized this modality as a ‘very neat and obvious use of interpreters’ (Paneth, 1957/2000, p.39, quoted by Braun (2015, p.353)) However, the first service for remote (telephone) interpreting was set up for bilateral mode operation in Australia in 1947 by Translating and Interpreting Service (TIS National), now under the Department of Home Affairs of the Australian Government, (Cabrera Méndez, 2016). Since then, remote interpreting services have increased year after year, particularly for bilateral interpreting, until reaching a peak during the COVID-19 pandemic.

Although technological improvements allow for more frequent and better communication, sound quality has always been a concern for telephone interpreters. Indeed, many researchers have focused on telephone interpreting technological needs emphasizing the importance of the sound quality of the phone and the connectivity of the line (Kelly, 2008), particularly in recent years when the use of mobile phones has increased. Similarly, videoconference interpreting poses comparable difficulties related to image quality and the position of cameras, which, if not used properly, might not register the necessary visual information.

Furthermore, some authors mention that remote interpreters suffer more stress than on-site interpreters (Andres & Falk, 2009), as well as an increased psychological effort and sense of alienation (Moser-Mercer, 2005, p.145) and augmented fatigue (Napier, Skinner & Braun,
2018). This is, in part, due to the extra effort required to adapt to new technologies at a late stage in their career and without the support of formal training.

Remote simultaneous interpreting was, at first, used as a solution for events held at venues without booths or where it was impossible to install portable booths, locating interpreters remotely elsewhere while the attendees continued within the conference room. Apart from this, in the conference interpreting technological sphere only the booths, receivers, infra-red radiators and interpreting consoles improved with new versions, and a simple change from analog to digital technology took place over many years in the profession.

However, COVID-19 caused the efforts devoted to improving remote simultaneous interpreting to skyrocket. Videoconferencing is now more and more present in our daily lives, and international events are using different platforms to connect people from all around the world. These platforms are slowly incorporating plugins that allow for the provision of interpretation, and improvements are made daily to ensure satisfactory results.

The global pandemic surprised professional conference interpreters, who were not used to investing in technology for their trade and did not have training in the use of such devices, beyond interpreter consoles in booths, which are used while working at multilingual events. As COVID-19 entered our lives, interpreting assignments were cancelled and the International Association of Conference Interpreters (AIIC) and the World Association of Sign Language Interpreters (WASLI), as well as the International Federation of Translators (FIT), issued a joint appeal to all governments requesting the inclusion of the majority of their freelance interpreters in financial rescue plans, since the organisers of national and international events were cancelling meetings and conferences across the board due to the health situation and mobility restrictions that said situation entailed (Runcieman, A. J., 2020). Faced with this scenario, conference interpreters found themselves at a crossroads: to wait until better times came when the health emergency had disappeared and a booth could once again be shared, or to accept the change and turn their offices into interpretation booths with their computers as interpreter consoles.

Although certainly many interpretation assignments were postponed and many others were cancelled "until further notice", the need for communication between speakers of different languages continued to persist: many international cooperation projects and projects with funding from international public institutions continued, and this required events to justify their progress. These events were held in the form of online webinars and they required remote simultaneous or consecutive interpretation by videoconference. Accepting the situation and adapting to the circumstances meant that professional interpreters, accustomed to in-person interpreting in the booth, had to become remote interpreters. But only a small percentage of these professionals accepted the challenge at the beginning, and both social networks and academic fora were filled with messages against it. We feel that one of the main reasons why many professional interpreters refused to become remote interpreters was due to the difficulty involved in changing the way of working for practising professionals without access to training for this change beforehand.

Despite the advantages that technology provides nowadays, professional remote interpreters struggle with the technical conditions, in terms of sound, connectivity and video, that make their job much more difficult than before the appearance of SARS-CoV-2. For this reason, we wanted to propose a contribution to the improvement of occupational health, both for current and future professional interpreters, by the designing a training module that includes a series of robust and specific technical recommendations, supported by active professional interpreters and telecommunication engineers, in terms of sound, hardware, connectivity and
the implications for computer performance or health, among other issues that must be included while training current and new professionals.

So we presented a part of our remote interpretation technology training course at the 43rd edition of the Translating and the Computer Conference (TC43) which ASLING (International Association for Advancement in Language Technology) held in November 2021. This course had been prepared for conference interpreters and was the result of a collaborative agreement between the University of Alcalá (Spain) and the telephone interpreting company Dualia Teletraducciones S.L. (Spain). This agreement has brought together both parties through a process of professional development and training for practising interpreters who mainly carry out their professional activity in public services, healthcare settings, and at international congresses.

The unit entitled ‘DEVELOPMENT OF TECHNOLOGICAL COMPETENCES: Remote Simultaneous Interpreting Explored’ is the one that was presented during the ASLING conference and whose main points are being included in this article to be recorded in the minutes and to serve as a knowledge transfer exercise for all professional colleagues who may benefit therefrom.

2 From onsite to offsite

There has been talk of remote interpretation since the publication by E. Paneth in 1957. In Spain, the government of the Basque Country financed the first startup that offered remote simultaneous interpretation for the nomination of San Sebastián as the Capital of Culture in 2016 (Sarasola & Aranberri, 2021), and in 2017 the first simultaneous online interpretation tasks were undertaken for statements by the then president, José Luís Rodríguez Zapatero (Jiménez Serrano, Ó. L., 2019). Nevertheless, these attempts were only the tip of the iceberg compared to the explosion that we have seen since mid-2020 with improvements in videoconferencing systems, the start of the Zoom Webinars platform as the first platform offering the possibility of having simultaneous remote interpretation at a very affordable price, better speed for communications, and a reduction in both bandwidth consumption and latency.

We have witnessed a major change in the way human beings participate in social events and now it is common to find events that are completely virtual or hybrid, in which the interpreters are located outside the physical venue and follow the conference from a remote platform for simultaneous interpretation which they access from their own computers. Through these platforms, the interpreters receive the audio and video signals from the speakers in real time and they return their interpretation exactly as if they were using an analog interpreter console. Participants can listen to the interpretation through the same platform or through a mobile application, both from the physical location of the event (if it is hybrid) and from anywhere else (if the event is completely online).

As with any change, remote simultaneous interpreting has many pros and a similar number of cons. Among the points in favour, there are the following:

• The service can be offered at 100% virtual events.

• There are cost savings that do not have to affect the interpreters’ fees, such as travel expenses, meal allowances, rental of physical equipment (portable booths, receivers, radiators, consoles, etc.).

• The platforms replicate the same functions as physical interpreter consoles and add enhancements in compliance with ISO accessibility standards.

• Carbon footprint reduction.
• It is possible to have interpretation services at events that otherwise could not have had these services due to a lack of local interpreters, budgetary issues, the pandemic, etc.

And among the points against, the following should be noted:

• There are a number of risks inherent in remote configurations, such as connection interruptions, reliability of the interpreters' electronic equipment, videoconferencing platform server downtime, etc.

• The sound quality of home equipment may not be as good as that of the professional equipment at traditional physical conferences.

• Lack of information and training on remote interpretation.

• Breaches in personal data security and confidentiality due to cloud-based systems.

• The additional pressure which interpreters are put under when the cognitive load increases while manipulating electronic devices and platforms at the same time.

3 Technical requirements

As was already mentioned, professional interpreters traditionally have not had the need to invest in physical equipment for their work, as they had everything necessary provided to them at the conference site. Thus, many questions arise when choosing the ideal headphones or microphone to offer the best level of service and to work with maximum comfort. We must start from the premise that the best microphone and headphones for one professional do not have to be the best for another professional, as there are ergonomic aspects and the professional's own physical preferences that make the choice something unique and personal. In addition, the performance of one device or another will also be affected by the rest of the equipment to which it is going to be connected: computer brand, operating system, browser, etc.

Recommendations for interpreters:

1. Computers should be connected to the internet by a wired connection, not by Wi-Fi, in order to guarantee the continuity of the connection throughout the entire work session.

2. We recommend using wired headphones instead of wireless headphones as, despite their comfort, wireless headphones need to be recharged and could lose power in the middle of an interpreting session.

3. Headphones can be connected to computers via a 3.5 mm jack or a USB port (Windows) or USB-C port (Mac). We recommend the latter option because, in this way, the graphics card is bypassed and no additional computer resources are consumed.

4. We recommend that the headphones be noise cancelling so that the circuit detects outside noise with built-in microphones and sends an equal but opposite cancellation signal to the headphones. By producing this countersignal, the headphones block a good deal of external sound sources.

5. Headphones with a built-in microphone must have a switch to turn the microphone OFF or ON. Turning the microphone ON and OFF with one's hands requires less cognitive load than with a keyboard shortcut or through mouse clicks in the platform interface.

6. Sound frequency is another piece of information that is not well known by interpreting professionals and it is necessary to remember that the human ear can perceive between 20 and 20,000 Hz, although the ISO standards for simultaneous interpreting (DIN EN ISO 20109 “Simultaneous Interpreting – Equipment – Requirements” and DIN EN ISO 20108 “Simultaneous Interpreting – Quality and Transmission of Sound and Image Input –
The ear can certainly receive above and below these values but it would be perceived as noise and we would not be able to decode it. According to experiments, as detailed by Yagura et al. (2021), this is the value that we must look for in our devices and the platform with which we are going to interpret.

7. The optimal headphone impedance should be between 16 and 32 ohms.

8. When choosing between a condenser microphone or a dynamic microphone, we recommend the second option because the condenser microphone has higher sensitivity, which will pick up unwanted background sounds, while the dynamic microphone only catches the sound coming from a single direction – from the interpreter's mouth.

9. We encourage professionals to have headphones with the microphone built in instead of handheld or tabletop microphones because an interpreter works with glossaries, jotting down data, looking at more than one screen, etc. All of this translates into continuous head movements. If the microphone is built into the earpiece and moves with the interpreter's head, it will always be the same distance from the mouth.

10. We encourage professional interpreters to invest in a UPS (Uninterruptible Power Supply). This is a battery that is plugged into a domestic power outlet, and the router and the computer can be connected to this battery so that, in the event of a power failure, the interpreter's equipment and connection would still operate for several hours.

But not everything depends on the interpreter; in fact, the source of the speech that the interpreter must translate comes from the speakers, who must also do their part to facilitate the interpreter's work.

The following minimum requirements should be asked of speakers:

1. Use an external microphone which is not the one built into the webcam. Microphones built into webcams are set for the equipment they are manufactured for and do not offer the sound quality that interpreters require.

2. Speakers must be aware that the "Share Screen" option robs them of prominence and, therefore, they must take more care in terms of their gestures – if they base their presentation on these.

3. Video playback uses up a large portion of bandwidth, so it is better to share videos once they have already been downloaded to the speaker's computer rather than directly from an online video platform. Otherwise, the video conferencing platform will adapt the bandwidth to be able to continue broadcasting and for the video to be displayed via the online platform without interruptions. In addition, speakers should remember that in many videoconferencing platforms it is necessary to activate the "share sound" option so that the rest of the participants can hear the audio of the video that is being projected.

4. Contrary to popular belief, natural light does a speaker a disservice by creating excess light, changing as the hours pass, and the speaker can even be left in unexpected darkness. It is preferable to have artificial light during the talk.

5. The frame, or the distance between the speaker and the camera, is something that is often forgotten about, yet distances should be well measured to facilitate gestures by the speaker. This is true whether the aim is to use the hands as an element of communication or to rely on face gestures.

6. Silence is a precious commodity in 100 % virtual events; thus, it is the responsibility of the participants to learn how to avoid excess noise by silencing their microphones when they
are not presenting and by controlling background noise. Of course, moderators should not allow speakers to present from noisy settings such as coffee shops or the street.

7. Webcam quality is often a common topic of discussion and equipment with 1080 pixels and 4k is highly appreciated; however, we must bear in mind that the resolution of the camera is not something that we should worry about a lot because it will be adapted to bandwidth constraints during the presentation. Thus, the real resolution will always be less than the maximum capacity that the camera can offer.

4 **Be aware of acoustic shock**

Acoustic shock is a phenomenon caused by the sudden explosion of a loud sound and the corresponding pressure spike that it generates. Its symptoms include nausea, vomiting, fainting, loss of balance, hearing loss, and ringing ears. Audio imbalances are common in video conferencing when there are several microphones that can be activated at the same time and all of them go directly to the ears of the interpreters. Therefore, the platforms used should include measures that reduce sound degradation as a result of an excess of active microphones.

5 **Something to take home**

Even (or perhaps above all) communication professionals who have achieved many successes at analogue events with large stages and a physical audience to interact with have to learn to handle these new environments if they want to generate the same impact as before. The fact is that both interpreters and speakers must adapt to the current times and learn to handle this situation, which has cropped up unexpectedly and in which communications technologies are constantly advancing. Lagging behind, after all, is not an option for anyone. These technologies mean that sound control is a new aspect that professionals must be prepared for; thus, the moment has come to invest in technical equipment to offer the best possible service. Having said that, equipment should always be bought bearing in mind that the best equipment for one interpreter in a booth may not necessarily be the best equipment for another.

**References**


DeepL vs Google Translate: which is the best at translating MWEs from French into Polish? A multidisciplinary approach to corpora creation and quality translation of MWEs

Emmanuelle Esperança-Rodier
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP*, LIG, 38000 Grenoble, France
emmanuelle.esperanca-roder@univ-grenoble-alpes.fr

Damian Frankowski
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP*, LIG, 38000 Grenoble, France
frankowskidmail@gmail.com

ABSTRACT
This article proposes a multidisciplinary approach to the creation of parallel French and Polish corpora annotated with multi-word expressions (MWEs) and the analysis of neural machine translation (NMT) errors in annotated MWEs, from French into Polish. The first task was to build the parallel FR-PL corpora from a French News corpus, taken from the WMT 2010 corpus of 40,000 words, by automatically translating it into Polish using the main commercial systems DeepL and Google Translate. The French source corpus had already been manually annotated with MWE, using the typology developed at LIDILEM (Tutin and Esperança-Rodier, 2019). In a second step, the quality of the MWE translations in the FR-PL parallel corpora was evaluated by annotating the translation errors after creating a new error typology based on MQM-DQF (Lommel et al., 2018a) and the linguistic features of MWE translations using the ACCOLE platform (Brunet-Manquat and Esperança-Rodier, 2018). Then, we selected 154 sentences (4,332 French words) from the MWE annotated French document and translated them into Polish using DeepL (3,599 Polish words) and Google Translate (3,519 Polish words). The first general result showed that for the French to Polish language pair, DeepL translated MWE better than Google Translate even though it used English as a pivot.

1. Introduction
Numerous studies and workshops have focused on multiword expressions (MWE), that can be defined as recurrent word combinations in which the general meaning does not correspond to the literal meaning of its individual lexical items (Firth 1957 and Sag et al., 2002). Some of these studies entailed the creation of several corpora as described in Constant et al. (2017). Amongst them we can cite two corpora of nominal and adverbial MWEs (Laporte et al., 2008a; Laporte et al., 2008b) which do not provide a typology. In addition, there is the French Treebank (Abeillé et al., 2003) which contains various MWEs, including verbal ones, but only in continuous expressions. As far as Polish is concerned, several MWE studies have been conducted, to mention just one, Savary (2001) presented the named entity annotation subtask of the Polish National Corpus. Finally, the ANR PARsing and Multiword Expression project (PARSEME - Project ANR-14-CERA-001), observing the lack of linguistic resources related to this topic, constituted a corpus of MWE syntactic annotations (especially verbal and nominal) (Candito et al., 2017) as well as tools for the analysis of MWEs.

MWE research is an active and topical research area as illustrated by the many conferences held, including the Multiword Expressions (MWE) workshops, organized since 2007 by the Special Interest Group on the Lexicon (SIGLEX) of the Association for Computational Linguistics (ACL) and supported by the Global Wordnet Association (GWA).

This article describes, the creation of parallel French and Polish corpora with annotated
MWEs, to subsequently analyze neural machine translation (NMT) errors in the annotated MWEs. We adopted the Tutin typology (Tutin and Esperança-Rodier, 2019), that describes a wide range of MWEs and allows the annotation of non-continuous expressions, as well as an NMT error typology adapted to MWE translations and based on the MQM-DQF typology (Lommel et al., 2018a).

After a short state-of-the-art presentation, we will address the methodology used by describing the MWE typology, the NMT error typology, the corpora and the evaluation tool. Then, we will present the results of the evaluation with examples and discuss these results to answer our question.

2. State of the Art

As Sag et al. (2002) reported, MWEs represent a real challenge for Natural Language Processing (NLP) and NMT which have to deal with this demanding task and may even reach record quality levels. Furthermore, Castilho et al. (2017) calculated that even if NMT systems can achieve brilliant results, the human evaluations were not as positive as the automatic metrics, especially as regards adequacy and post-editing effort. Koehn and Knowles (2017) also demonstrated that although NMT has achieved some success, it still faces various challenges, notably out-of-domain performance and limited resources.

Since entire sentences are converted to vectorial representations in NMTs (Rikters and Bojar, 2017), and due to the absence of phrasal segmentation in NMTs (Zaninello and Birch, 2020), MWEs are particularly difficult to identify. Colson (2020) reports that in about 40% of MWE translations Google Translate made a mistake.

Automatic metrics are often difficult to interpret and do not identify the main translation problems as Vilar et al (2006) found and proposed an NMT error typology in order to facilitate the analysis of the translation errors. In the literature, error typologies are used to assess NMT outputs as Popović (2018) mentioned and Lommel (2018b) proposed a standardization of the error classification with MQM-DQF.

The study we describe below, positions itself amongst these efforts, proposing a way of annotating all the MWE types, both continuous and non-continuous, and proposing a standardized translation error typology adapted to MWEs.

3. Methodology

The first step was to build two parallel FR-PL corpora, taking the French News WMT 2010 corpus (40,000 words), and automatically translating it into Polish using DeepL and Google Translate. The French corpus already had manually annotated MWEs, according to the typology developed at LIDILEM (Tutin and Esperança-Rodier, 2019). Then, the quality of the MWE translations in the FR-PL parallel corpora were evaluated by annotating the translation errors and the linguistic features of these MWE translations using the ACCOLÉ annotation platform (Brunet-Manquat and Esperança-Rodier, 2018).

Descriptions are given below of the corpus creation, the MWE typologies, the MT Error typologies used, and ACCOLÉ.

3.1. Corpus

We used the WMT 2010 News corpus, as it had already been annotated according to the MWE typology described below. To start, we selected the first 154 sentences, representing 4,332 French words, from the French document and translated them into Polish using Google Translate and then DeepL, thus obtaining two translated documents of respectively 3,519 and 3,599 Polish words, as illustrated in figure 1.
3.2. MWE Typology

We adopted the Tutin typology (Tutin et al., 2019), as described in Table 1, as it allows the annotation of non-continuous expressions and covers a large range of MWEs. Nevertheless, it should be noted that the typology consists of eight types, including routine formulae, such as “it must be noted” and Pragmatic MWEs “you’re welcome”.

<table>
<thead>
<tr>
<th>Multiword expressions</th>
<th>Descriptions</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idioms</td>
<td>frozen multiword expressions</td>
<td>cul de sac (fr) ‘dead end’; prendre en compte (fr) ‘take into account’</td>
</tr>
<tr>
<td>Collocations</td>
<td>preferred binary association, including light verb constructions</td>
<td>gros fumeur (fr) ‘heavy smoker’; faire une promenade (fr) ‘to take a walk’</td>
</tr>
<tr>
<td>Functional Multiword Expressions</td>
<td>functional adverbs, prepositions, conjunctions, determiners, pronouns.</td>
<td>c’est pourquoi (fr) ‘that is why’; d’autre part (fr) ‘on the other hand’; insofar as</td>
</tr>
<tr>
<td>Pragmatic MWEs</td>
<td>multiword expressions related to specific speech situations.</td>
<td>de rien (fr) ‘You’re welcome’; à plus tard (fr) ‘see you later’.</td>
</tr>
<tr>
<td>Proverbs</td>
<td></td>
<td>Pierre qui roule n’amasse pas mousse (fr) ‘A rolling stone gathers no moss’</td>
</tr>
<tr>
<td>Complex terms</td>
<td></td>
<td>natural language processing</td>
</tr>
<tr>
<td>Multiword Named entities</td>
<td></td>
<td>Université Grenoble Alpes; the European Union;</td>
</tr>
<tr>
<td>Routine formulae</td>
<td>routines generally associated to rhetorical functions</td>
<td>force est de constater (fr) ‘it must be noted’.</td>
</tr>
</tbody>
</table>

Table 9: Tutin et al (2019) MWE typology

In our typology, it should also be noted that idioms include compound nouns; moreover, only multiword named entities are considered as MWEs. As well as annotating the type of the MWE, the Part of Speech (POS) of the MWE was also annotated.
Once translated, a qualitative evaluation was made of the two Polish documents using the collaborative Platform ACCOLÉ, described below.

### 3.3. Annotation Platform: ACCOLÉ

ACCOLÉ is a collaborative error annotation platform described by Brunet-Manquat et al. (2018). To annotate the errors made by the two NMT systems, we created two projects, one for each NMT system.

As shown in Figure 2, ACCOLÉ displays the French source and the Polish translation. The platform indicates a potential MWE by underlining several words in the French source. The annotator, a native Polish linguist graduate in English studies at Jagiellonian University and French studies (translation and interpretation) at Grenoble Alps University, selected the French MWEs as well as their Polish translations and annotated them according to the NMT error typology described below.

### 3.4. NMT Error Typology

We elaborated our own typology based on the MQM-DQF (Lommel et al, 2018a) addressing NMT errors, as MQM-DQF aims to be a harmonized typology, and offers a wide range of error types best suited to the Polish language; but also, due to the way MWEs were translated, as shown in Figure 3. “Different POS” refers to an MWE where the part of speech

![Figure 2: ACCOLÉ screenshot of MWE and NMT error annotation](image)

![Figure 3: MWE annotation](image)
was different from the original French one. “Different Type” means that the MWE translated into Polish was of a different type than the French one. “MWE translated” indicates the case when the French MWE was translated into a Polish MWE as well. “Term translated” denotes the translation of a French MWE into Polish as a single term. We added “identification” because of an issue with the internal identification. “Literal” is the label which we gave to the literal translation of an MWE into Polish.

Inspired by MQM-DQF metrics, we assigned a score to each MWE translation. Yet, we did not compute any metrics. Score 0 was assigned when the MWE was translated incorrectly, with a mistake; Score 1 when the MWE translation was comprehensible, without a mistake, although it could be corrected to make the translation more appropriate to the context; and finally for translations that were well translated and without a mistake, a score of 2 was attributed.

Turning to the MT errors, we selected the items provided by MQM-DQF typology (Lommel et al, 2018a), listed in Figure 4.

We selected four types for Accuracy. When a MWE is added in the Polish translation, the type is “Addition”, and when a MWE is missing in the translation the type is “Omission”. When the French MWE is not translated into Polish, the type is “Untranslated”, and when the French MWE is not translated correctly the type is “Mistranslated”.

For “Fluency”, six types were selected, three of which, “Spelling”, “Style” and “Typo”, were divided into subtypes.

We tested our typology on fifty sentences to check that all the cases encountered could be annotated using our typology.

After describing the MWE typology, and the MT error typology used to annotate the Google Translate and DeepL translation quality within ACCOLE, a description is now given of the evaluation results.

4. Evaluation Results

In order to provide a better picture of our data, we performed a short analysis of the MWEs contained in the French document.

4.1. French-Source MWE-distribution

As shown previously in Figure 1, the French source corpus consisted of 4,332 words. We
annotated 1,546 words as being MWEs, which represented 35% of the total words. This figure complies with Cartier (2008) who showed that MWEs represented 30% of the units in a given document. Among these 1,546 words, we mainly annotated “Collocations” (32%), and “Function Words” (“Functional Multiword expressions” in the MWE typology) (24%), as shown in Figure 5. We had only two “Pragmatemes” (“Pragmatic MWEs”) representing 0% in our data and no proverbs.

![Figure 5: MWE distribution in French Source Document](image)

But we did have some multiword “Named Entities” (18%), and some “Full Phrasemes” (“idioms” - 15%). Then, to a lesser extent, we had a few “Complex Terms” (7%), and “Routine Formulae” (4%).

Keeping this in mind, we now provide the quality evaluation results for Google Translate and DeepL.

### 4.2. Google Translate hypothesis evaluation results

Figure 6 shows that most of the translation errors were made in “Collocations” (C - 34%), then, in order of occurrence, in “Function words” (F - 20.5%), “Full Phrasemes” (PH - 16%), “Complex Terms” (T - 13%), “Routine Formulae” (R - 8%), “Named Entities” (NE - 8%) and a negligible number in “Pragmatemes” (P - 0.5%). As regards the distribution of MWEs in the source document, it is logical that most of the translation errors are found in the Collocation type as this is the most numerous. The same applies for “Function Words”. On the other hand, we could not conclude anything about “Pragmatemes” as they were under-represented in our corpus, so we do not analyze this type below. Nevertheless, it should be noted that Google Translate made two mistakes in this MWE Type.
As shown in Appendix A, the very few errors made with “Named Entities” appear in all the “Accuracy” error types while they occur in seven of the ten “Fluency” error types. “Mistranslation” is the main error type for “Routine Formulae”, while for the “Fluency” error types, “Grammar” is the main error followed by “Unintelligible” and the two “Style” types. “Complex Terms” show the same trend as for “Accuracy” errors, but as regards the “Fluency” errors, “Unintelligible” comes second, followed by “Style-unidiomatic” and “Grammar” respectively. Looking at “Function Words”, most of the “Accuracy” errors were of the “Mistranslation” type, with the “Omission” errors ranking second. Still in “Function Words”, most of the “Fluency” errors were the “Unintelligible” type, followed by “Style-unidiomatic”, “Style”, “Spelling” and “Capitalization”. To conclude with MWE types, “Collocations” being the most represented MWE type in the source document, accounted for most of the “Accuracy” error types, mainly “Mistranslation” errors. “Collocations” also had the most “Fluency” errors, of the “Style-unidiomatic” type, followed by “Unintelligible”, “Style”, “Grammar”, “Duplication” and “Capitalization”.

Translation errors were mainly, almost 66%, “Fluency” errors, with “Unintelligible” ranking first at 20.5% of the total errors and roughly 31% of “Fluency” errors, occurring in the whole range of MWE types, but principally in “Collocations” and “Full Phrasemes”. The “Unintelligible” type was followed by 17% of the “Style-unidiomatic” type, a quarter of the “Fluency” errors and found through all the MWE types but mainly in “Collocations”. The third most common error type was “Style”, representing 9% of the total errors, and 14% of the “Fluency” errors, again found in the whole range of MWE types but mainly in “Collocations”, then “Full Phrasemes” and “Function terms”. Finally, the “Grammar” type occurred to a lesser extent (11% of total errors but 17% of “Fluency” errors) and almost equally among all the MWE types, with a small peak in “Collocations”.

“Accuracy” errors, 34% of total errors, with “Mistranslation” the most frequent of the MWE types (26% of total errors, and 77% of “Accuracy” errors), occurring over the whole range of MWE types, as well as “Omission” (16% of “Accuracy” errors and merely 5.5% of total errors), also spread over all the MWE types but to a much lesser extent. Finally, the “Addition” type (just 3% of “Accuracy” errors), which only occurred in “Function terms” and “Named Entities”, and the “Untranslated” type (4% of “Accuracy” errors), found in “Full phrasemes” and “Named Entities.”
Table 2: Error ranking according to MWE types

<table>
<thead>
<tr>
<th>MWE type</th>
<th>#Words/types</th>
<th>% of total Words</th>
<th>% of MWE Words</th>
<th>MWE type Rank</th>
<th>#Total Errors/MWE</th>
<th>% Error</th>
<th>Error Rank</th>
<th>Google Translate</th>
<th>#Total Errors/MWE</th>
<th>% Error</th>
<th>Error Rank</th>
<th>DeepL</th>
</tr>
</thead>
<tbody>
<tr>
<td>collocation</td>
<td>498</td>
<td>2.30</td>
<td>32.21</td>
<td>1</td>
<td>70.00</td>
<td>33.98</td>
<td>1</td>
<td>20</td>
<td>48.78</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>function word</td>
<td>377</td>
<td>1.74</td>
<td>24.39</td>
<td>2</td>
<td>33.00</td>
<td>16.02</td>
<td>3</td>
<td>10</td>
<td>24.39</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>named entity</td>
<td>279</td>
<td>1.29</td>
<td>18.05</td>
<td>3</td>
<td>16.00</td>
<td>7.77</td>
<td>6</td>
<td>6</td>
<td>14.63</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>full phraseme</td>
<td>231</td>
<td>1.07</td>
<td>14.94</td>
<td>4</td>
<td>42.00</td>
<td>20.39</td>
<td>2</td>
<td>10</td>
<td>24.39</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>complex term</td>
<td>100</td>
<td>0.46</td>
<td>6.47</td>
<td>5</td>
<td>27.00</td>
<td>13.11</td>
<td>4</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>routine formulae</td>
<td>59</td>
<td>0.27</td>
<td>3.82</td>
<td>6</td>
<td>17.00</td>
<td>0.08</td>
<td>5</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>pragmateme</td>
<td>2</td>
<td>0.01</td>
<td>0.13</td>
<td>7</td>
<td>1.00</td>
<td>0.49</td>
<td>7</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>1546</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>206</td>
<td>-</td>
<td>41</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

As can be seen in Table 2, the second-highest error ranking corresponds to “Full Phrasemes” that only represents 15% of all the MWE types. “Full Phrasemes”, however, ranked fourth in the MWE type occurring in the French source document. Unlike “Named Entities”, which ranked third in the MWE types in the French document, but only sixth for errors. Google Translate also had more difficulties with “Fluency” than with “Accuracy”, and the type of errors, which were, respectively, mostly “Unintelligible” and “Mistranslations”.

It is possible to draw the conclusion that Google Translate had difficulty in translating “Full Phrasemes”, the fourth most represented MWE type in the French text but ranked second for the most errors, while it performed quite well with “Named Entities”, ranked third in the MWE types in the French document, but only sixth for errors. Google Translate also had more difficulties with “Fluency” than with “Accuracy”, and the type of errors, which were, respectively, mostly “Unintelligible” and “Mistranslations”.

Examining the examples given in Appendix B, taken from the research in this study, we will first discuss the examples of the “Mistranslation” type taken from Google Translate.

In [1], the French term Gouverneur ‘Governor’ was translated by Google Translate into Polish as gubernator which is a literal translation. It should however be translated as ‘president of the national bank of Poland’, which in Polish is rendered with prezes. The word gubernator, in the Polish language is used to refer to the governor of the state in the USA, Connecticut or Illinois for example.

In [2], the expression faire la guerre ‘to wage war’ was not translated correctly into Polish. The exact translation should be prowadzić wojnę meaning in French faire la guerre. Google Translate rendered the French expression faire la guerre as a wyruszyć na wojnę ‘to set off’, ‘to go to war’.

The next example [3] sounds a little bit amusing and improbable. À la seule condition que ‘with the only condition that’ was translated into Polish using the word podeszwy, ‘soles’, which means the sole of one’s foot and has nothing to do with the context. The correct version in Polish is pod jednym warunkiem, że.

In [4], the phrase pour que is used to express purpose in French. Nevertheless, Google Translate has rendered it as a conjunction żeby ‘that’. The mistake may result from the fact that in Polish a very similar word żeby is used to express purpose.
Then there are the “Unintelligible” translations. Looking at [5], this fifth example is also absurd and amusing.

In [6], there is the word itonia. The correct translation is the country named Estonia ‘Estonia’, but the letters e and s were omitted.

In [7], the words l’adoption de l’euro ‘the euro adoption’ were repeated several times and there is also an apostrophe which is not required in this context. This is strange as stammering happens only with MT systems which are not well trained. And Google Translate should be well trained.

4.3. DeepL hypothesis evaluation results

Now for an analysis of the results obtained while performing the Quality Evaluation of the DeepL MWE translations.

Figure 7 shows that most of the translation errors were made in “Collocations” (C-49% of the total errors) followed by “Function Terms” (T-24%), “Named Entities” (NE-15%) and finally “Full Phrasemes” (PH-12%). DeepL did not make any mistakes with “Complex Terms”, “Routine Formulae” or “Pragmatemes”. The distribution of MWE types over the source document did not provide sufficient occurrences of “Pragmatemes” to conclude anything. However, in the two words annotated as “Pragmateme” MWE type, DeepL did not make any mistakes.

![Figure 7: DeepL error type per MWE type](image)

As shown in Appendix C, “Collocations” accounted for most of the Translation errors, especially the “Fluency” type, mainly in the two “Style” error types, followed by the “Unintelligible” and “Capitalization” and “Grammar” types. while it was only the second MWE type for the “Accuracy” errors. The highest “Accuracy” error ranked was in the “Function Term” MWE type, with all the “Accuracy” mistakes labeled as the “Omission” type, while no error was labeled under “Fluency”. The third MWE type with the most errors was the “Named Entity” type, with four errors in “Accuracy”, mostly of the “Untranslated” type, and only two “Fluency” errors i.e. “Unintelligible” and “Duplicated”.

Looking at the “Accuracy” errors which represent most of the errors made by DeepL with 58.5% of the total errors, most of those errors were of the “Omission” type, with 34% of the total errors and 58% of the “Accuracy” errors. “Omissions” occurred most in “Function Words”, then to a much lesser extent in “Full Phrasemes”. “Mistranslations” ranked second,
representing 15% of the total errors and 25% of the “Accuracy” errors, occurring only in “Collocations” and “Named entities” (one occurrence). The “Untranslated” error type came third, representing 7.5% of the total errors and 12.5% of the “Accuracy” errors, mainly in “Named Entities”. “Addition” was the last error type made by DeepL (2.5% of total errors and 4% of “Accuracy” errors) and with only one error in “Named Entities”.

Nevertheless, as can be seen in Table 2, leaving “Collocations” in first place, “Full Phrasemes” rank second in the translation error type scale, equal to the “Function” Word type, while representing 15% of the MWE types in the French source document.

This suggests that DeepL has difficulty in translating “Full Phrasemes”, while it performs pretty well with “Complex Terms” and “Routine Formulae”. When DeepL makes errors, they are mainly of the “Accuracy” type, usually “Omission”. As for the “Fluency” type, DeepL struggles with the “Style”.

In Appendix B example [8], DeepL has translated the whole sentence into Polish yet the expression le Projet d’assistance à l’école secondaire féminine was only translated into English. This may be evidence of the use of English as a pivot language by DeepL.

To illustrate the “Omission” error type, [9] shows that the omission of de plus en plus ‘more and more’ implied a Score 0. Although the phrase de plus en plus was not translated into Polish, the meaning of the translated sentence was not affected and the accuracy as well the fluency was preserved. If we wish to translate the phrase de plus en plus, the translation would be coraz to bardziej.

**4.4. MWE translation evaluation**

After investigating the MT errors in both Google Translate and DeepL translations, it is possible to analyse the way MWEs have been translated.

Both DeepL and Google Translate correctly translated many French MWEs into equivalent Polish MWEs. Google Translate translated a few more French MWEs into terms than DeepL but did double the number of literal translations. Finally, DeepL more frequently used a Polish MWE with a different POS and a different MWE type than Google Translate.

<table>
<thead>
<tr>
<th>MWE translated</th>
<th>Google Translate</th>
<th>DeepL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term translated</td>
<td>163</td>
<td>150</td>
</tr>
<tr>
<td>Literal translation</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td>Different POS</td>
<td>54</td>
<td>77</td>
</tr>
<tr>
<td>Different type</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 4: Number of MWEs translated by Google Translate and DeepL according to the different MWE types.

**4.5. Score evaluation**

Looking at the scores (“Score 0” for incorrect translations, “Score 1” for approximate translations which could be improved, and “Score 2” for correct translations) it is possible to see that Google Translate produces many correct translations but, makes more incorrect and approximate translations than DeepL. Furthermore, DeepL makes far fewer mistakes and provides more correct translations.
A comparison can be made of the results of Google Translate and DeepL by analysing the NMT errors, the MWEs and the Scores.

DeepL makes far fewer errors than Google Translate (cf. Table 2), respectively 41 error annotations versus 206, which is confirmed by the scores. Google Translate makes errors across the whole range of MWE types, while DeepL only makes mistakes in “Collocations”, “Function Words”, “Named Entities” and “Full Phrasemes”. DeepL makes more “Accuracy” errors than Google Translate (respectively 58.5% and 34%) which, in turn, makes more “Fluency” errors (respectively 41.5% and 66%). Google Translate produces more “Mistranslation” and “Unintelligible” translations than DeepL which makes more “Omission” and “Style” errors.

DeepL uses more MWEs with a different POS or MWE type than Google Translate which translates more literally and uses more terms instead of equivalent Polish MWEs (cf. Table 3).

Referring to Appendix B, example [10] shows that both Google Translate and DeepL translate in exactly the same way and make a “Style-unidiomatic” error. In the phrase *offrir une deuxième chance*, the verb *offrir* ‘to offer’ was translated into Polish literally as *zaoferować*. Nevertheless, the expression *zaoferować drugą szansę* is unidiomatic and should be *dać* which also means *offrir*.

In example [11] the noun *l’offre* ‘the offer’ derived from the verb *offrir* already seen in example [10]. In this case, the noun was translated correctly by Google Translate but not by DeepL.

In example [12], the collocation *dure réalité* ‘harsh reality’ has been mistranslated in both cases. The Polish translation is correct but it is inappropriate in this context. Instead of using the adjective *ostrej* ‘sharp’ in the Google ‘Translate hypothesis or *twardej* ‘hard’ in the DeepL hypothesis, in Polish it would be preferable to say *trudna* ‘difficult’ which would appear more natural.

In the next example [13] of “Style” error type, both systems have translated in a way that does not accurately represent the MWE in the source sentence. The sentence rendered by Google Translate is comprehensible and grammatically correct. Nonetheless, the Polish sentence has a stylistic error. Studying the MWE *ułatwić dostęp do edukacji na poziomie średnim* we believe that this translation needs to be rephrased. As such, it would be better to say *ułatwić dostęp do edukacji w szkole średniej* ‘facilitate access to education at secondary school’.

In example [14], *premier cycle d’éducation* should also be corrected to sound more natural and less artificial in Polish. Therefore, instead of *szkołę średnią I stopnia* it would be better to say *pierwszy etap edukacji* ‘the first stage of education’.
6. Conclusion

After considering the NMT errors, MWE translations, and Score evaluations, as well as the different examples, it is possible to answer our question, namely “DeepL vs Google Translate: which is the best at translating MWEs from French into Polish?”. Clearly DeepL translates French MWEs into Polish MWEs with far fewer errors than Google Translate, even if it has a tendency for “omissions” and struggles with the “style” error types. DeepL also uses more MWEs with different POS or MWE types than the ones in the French source. This, as translation studies have already demonstrated, denotes that it does not impact the translation quality. Both NMT systems struggled to translate Full Phrasemes. We also found that DeepL used English as a pivot between French and Polish. Finally, as the length of the sentences was homogeneous, it was not possible to see whether the length of a sentence affected any specific mistakes.

Continuing this study, we will complete the annotation of the FR-PL translation corpus. We will also investigate the causes of the NMT errors. We plan to study, using this corpus, the different factors that influence the translation of MWEs as MWEs, Terms or Literal translations, and which factors are involved in determining whether MWEs are more likely to be translated with different POS or MWE types.

Acknowledgements

We would like to thank Pôle Grenoble Cognition which funded the internship during which this work was performed.

References


Lommel, Arle, and Alan K. Melby (2018a) Tutorial: MQM-DQF: A Good Marriage (Translation Quality for the


Appendix A: Google Translate - Number of Translation errors per MWE types. The same MWE can be annotated with several translation error types.

<table>
<thead>
<tr>
<th>Error Types/MWE</th>
<th>C</th>
<th>PH</th>
<th>F</th>
<th>T</th>
<th>R</th>
<th>NE</th>
<th>PRAG</th>
<th>TOTAL</th>
<th>TOTAL (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY (TOTAL)</td>
<td>18</td>
<td>13</td>
<td>16</td>
<td>11</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>70</td>
<td>33.98</td>
</tr>
<tr>
<td>addition</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0.97</td>
</tr>
<tr>
<td>mistranslation</td>
<td>17</td>
<td>9</td>
<td>11</td>
<td>10</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>54</td>
<td>26.21</td>
</tr>
<tr>
<td>omission</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>5.34</td>
</tr>
<tr>
<td>untranslated</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1.46</td>
</tr>
<tr>
<td>FLUENCY (TOTAL)</td>
<td>52</td>
<td>29</td>
<td>17</td>
<td>16</td>
<td>10</td>
<td>12</td>
<td>0</td>
<td>136</td>
<td>66.02</td>
</tr>
<tr>
<td>duplication</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>2.91</td>
</tr>
<tr>
<td>grammar</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>23</td>
<td>11.17</td>
</tr>
<tr>
<td>spelling</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>2.43</td>
</tr>
<tr>
<td>capitalization</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>2.43</td>
</tr>
<tr>
<td>diacritics</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>style</td>
<td>9</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>19</td>
<td>9.22</td>
</tr>
<tr>
<td>style - unidiomatic</td>
<td>20</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>34</td>
<td>16.50</td>
</tr>
<tr>
<td>punctuation</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.49</td>
</tr>
<tr>
<td>typo space</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.49</td>
</tr>
<tr>
<td>unintelligible</td>
<td>12</td>
<td>10</td>
<td>7</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>42</td>
<td>20.39</td>
</tr>
<tr>
<td>TOTAL</td>
<td>70</td>
<td>42</td>
<td>33</td>
<td>27</td>
<td>17</td>
<td>16</td>
<td>1</td>
<td>206</td>
<td>-</td>
</tr>
<tr>
<td>TOTAL (%)</td>
<td>33.98</td>
<td>20.39</td>
<td>16.02</td>
<td>13.11</td>
<td>8.25</td>
<td>7.77</td>
<td>0.49</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
# Appendix B: Translation Error Examples

## MISTRANSLATION

Google Translate

[1] **FR:** Comme le fait remarquer Leszek Balcerowicz, le Gouverneur de la Banque nationale de Pologne, l'adoption de l'euro par les pays candidats [...]  
*PL Hyp.:* Ponieważ latka Balcerowicz wskazuje, że gubernator Narodowego Banku Polskiego, przyjęcie euro przez kraje kandydujące [...]  

[2] **FR:** [...] plus personne n'estimera utile de faire la guerre pour déterminer quelle langue doit être parlée en Alsace-Lorraine.  
*PL Hyp.:* [...] aby nikt nie uznał za pożyteczne wyruszenia na wojnę w celu ustalenia, jakim językiem należy się posługiwać w Alzacji i Lotaryngii.  

[3] **FR:** Mais contre toute attente, il a adopté une position éthiquement avantageuse en proposant de renoncer à une partie de son rabais budgétaire, à la seule condition que le reste de l'Union européenne (et la France, en particulier) s'accorde sur les réformes fondamentales en matière de politique agricole européenne.  
*PL Hyp.:* Jednak wbrew wszelkim szansom, przyjęło etycznie korzystne stanowisko, proponując rezygnację z jej zniżki budżetowej, podeszwy, że reszta Unii Europejskiej (i w szczególności w szczególności) zgadza się na reformy fundamentalne w obszarze Europejska polityka rolna.  

[4] **FR:** Il a fallu du temps pour que les Britanniques comprennent qu'étant donné que le Président Jacques Chirac [...]  
*PL Hyp.:* Brytyjczycy zajęło to, że Brytyjscy zrozumieli, że ponieważ prezydent Jacques Chirac [...]  

## UNINTELLIGIBLE

Google Translate

[5] **FR:** Ils pourraient commencer par considérer l'adoption précoce de l'euro par les pays candidats d'un œil plus favorable, aussi bien pour les pays disposant d'un système de caisse d'émission que pour ceux utilisant un taux de change flottant  
*PL Hyp.:* Mogliby zacząć od rozważenia wczesnego przyjęcia euro przez kandydujące kraje bardziej korzystnego oka, zarówno dla krajów z systemem ciasta, jak i dla osób korzystających z pływającego kursu.  

[6] **FR:** Trois d'entre eux --l'Estonie, la Lettonie et la Lituanie-- possèdent un système de caisse d'émission ou un taux de change fixe avec l'euro, tout comme la Bulgarie, qui doit rejoindre l'Union en 2007  
*PL Hyp.:* Trzy z nich –itonia, Łotwa i Litwa mają system funduszy emisji lub stałego kursu z euro, podobnie jak Bułgaria, aby dołączyć do Unii w 2007 roku.  

[7] **FR:** Malheureusement, l'entêtement de la BCE et de la Commission apporte maintenant de l'eau au moulin des gouvernements populistes dans les pays accédants qui comptent sur la variation des cours de change pour échapper à la discipline fiscale qu'implique l'adoption de l'euro  
*PL Hyp.:* Niestety, upór EBC i Komisja przynoszą teraz wodę do młynów rządów populistycznych w krajach przystępujących, które polegają na zmianie kursów walutowych, aby uciec przed dyscypliną podatkową dorzuconą przez przyjęcie przyjęcia przyjęcia Przyjęcie przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia przyjęcia z'euro
Au Bangladesh, le Projet d’assistance à l’école secondaire féminine a aidé financièrement les jeunes filles âgées de 11 à 14 ans à poursuivre leurs études et à retarder le moment du mariage.

*PL hyp.: W Bangladeszu projekt Women's Secondary School Assistance Project zapewniał pomoc finansową dziewczicom w wieku od 11 do 14 lat, aby mogły kontynuować naukę i opóźnić zawarcie małżeństwa.

Les gouvernements doivent aussi offrir une 'deuxième chance' aux jeunes qui ont échoué en raison des circonstances ou de mauvais choix.

*PL Hyp.: Rządy muszą również zaoferować "drugą szansę" młodym ludziom, którzy zawiodły okoliczności lub złe wybory.

Une fois que les Britanniques ont pris conscience de cette dure réalité, ils ont joué leur lamentable dernière carte dans la gestion des négociations.

*PL Hyp.: Kiedy Brytyjczycy zdali sobie sprawę z tej twardzej rzeczywistości, zagrali swoją żałosną ostatnią kartę w zarządzaniu negocjacjami.
Plusieurs pays essayent maintenant de faciliter l'accès à l'enseignement secondaire, notamment grâce à des programmes conditionnels de transferts [...]  

*PL Hyp.: Obecnie kilka krajów próbuje ułatwić dostęp do edukacji na poziomie średnim, w szkole średniej w szczególności poprzez programy warunkowego przeniesienia, „[…]  

PL: Obecnie kilka krajów próbuje ułatwić dostęp do edukacji, w szkole średniej w szczególności poprzez programy warunkowego przeniesienia, „[…]  

Au Maroc par exemple, plus de 80% des enfants accomplissent un premier cycle d'éducation, mais moins de 20% atteignent le niveau requis.  

*PL Hyp.: W Maroku, na przykład, ponad 80 procent dzieci kończy szkołę średnią I stopnia, ale mniej niż 20 procent osiąga wymagany poziom.  

PL: W Maroku, na przykład, ponad 80 procent dzieci kończy pierwszy etap edukacji, ale mniej niż 20 procent osiąga wymagany poziom.
Appendix C: DeepL - Number of Translation errors per MWE type. The same MWE can be annotated with several translation error types.

<table>
<thead>
<tr>
<th>Error Types/MWE</th>
<th>C</th>
<th>PH</th>
<th>F</th>
<th>T</th>
<th>R</th>
<th>NE</th>
<th>PRAG</th>
<th>TOTAL</th>
<th>TOTAL (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCURACY</td>
<td>7</td>
<td>3</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>24</td>
<td>58.54</td>
</tr>
<tr>
<td>addition</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2.44</td>
<td></td>
</tr>
<tr>
<td>mistranslation</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>14.63</td>
<td></td>
</tr>
<tr>
<td>omission</td>
<td>1</td>
<td>3</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>34.15</td>
<td></td>
</tr>
<tr>
<td>untranslated</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>7.32</td>
<td></td>
</tr>
<tr>
<td>FLUENCY</td>
<td>13</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>17</td>
<td>41.46</td>
</tr>
<tr>
<td>duplication</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2.44</td>
<td></td>
</tr>
<tr>
<td>grammar</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>4.88</td>
<td></td>
</tr>
<tr>
<td>spelling</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>capitalization</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2.44</td>
<td></td>
</tr>
<tr>
<td>diacritics</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>style</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>12.20</td>
<td></td>
</tr>
<tr>
<td>style - unidiomatic</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>12.20</td>
<td></td>
</tr>
<tr>
<td>punctuation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>typo space</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>unintelligible</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>7.32</td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>20</td>
<td>5</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>41</td>
<td>48.78</td>
</tr>
<tr>
<td>TOTAL (%)</td>
<td>48.78</td>
<td>12.20</td>
<td>24.39</td>
<td>0.00</td>
<td>0.00</td>
<td>14.63</td>
<td>0.00</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Introducing the PETIMOD Corpus: A Resource for the Analysis of Personification in EU Mediated and Non-Mediated Discourse

Fernando Sánchez Rodas

University of Malaga
fersanchez@uma.es

Abstract

Personification is one of the most basic ontological metaphors, a linguistic representation of a mental mechanism which pervades human cognition (Lakoff and Johnson, 1980; MacKay, 1986). However, the relation between the linguistic forms and the conceptual structures of personification has not been discussed fully and systematically yet (Dorst, 2011), let alone in translated and interpreted discourse. This paper will try to look at a specific type of humanising metaphor, called personification-with-metonymy (Dorst et al., 2011), in PETIMOD, an English-Spanish intermodal corpus of the European Parliament Committee on Petitions (Corpas Pastor and Sánchez Rodas, 2021/In press; Corpas Pastor and Sánchez Rodas, 2022/In press). The methodology used is innovative and combines the MIPVU procedure (Steen et al., 2010) with the automatic extraction of $N+V$ patterns with named entities by means of VIP (Corpas Pastor, 2020). Results suggest the simplification of ENT(=ORG) + $V$ patterns in mediated discourse, but not of personifications, which are indeed more numerous in interpreted speeches (67.5 % vs. 71.4 %). Interpreted discourse also shows more personifications of speech and action than of reasoning, which could be a consequence of orality. We conclude that a constructional approach may be necessary for further study of the metonymy-metaphor relations in personification.

1 Background

According to Lakoff and Johnson (1980, 33), personification is an ontological metaphor involving a cross-domain mapping where an object or entity “is further specified as being a person.” Since people use human beings as their prototypical or default frame of reference, personification pervades human cognition and often comes disguised in other figurative devices or syntactic expressions (MacKay, 1986). However, the relation between the linguistic forms and the conceptual structures of personification has not been discussed systematically neither in the field of Linguistics nor in Natural Language Processing (NLP), and the influence of factors such as conventionality, deliberateness and metonymy has not received much attention in the past decades (Dorst, 2011: 132). In an empirical study, Dorst et al. (2011) offered an integrated typology for the classification of personifications in discourse. This study included four categories: conventionalized personification, novel personification, default personification, and personification-with-metonymy, i.e., metonymical personifications. Metonymical personifications are based on a violation of the selection restrictions of the basic sense caused by the replacement of a human agent or patient with a metonymically related non-human agent or patient, e.g. The CND office telephoned to ask her for voluntary evening help (Dorst et al., 2011: 148). This type of personification is underlined as a potentially interesting area for further research; although a metonymic relation was involved, the participants of the study were still able to see the possibility of a personification interpretation (ibid.: 194). Besides, literature examples of metonymical non-human agents or patients in these personifications often include names, such as Paris and Washington are having a spat (Evans, 2007: 143) or The White House denies… (Dorst, 2011: 132). The reasons behind the choice of names can be deemed pragmatic (e.g., as a mechanism to avoid responsibility in journalism), but also of semantic accessibility (Croft, 1993).
2 Hypothesis and objectives

In Corpas Pastor and Sánchez Rodas (2022/In press), a sample of verbal collocations with the names Commission/European Commission was extracted from an English-Spanish corpus of mediated and non-mediated Eurolect. Their analysis revealed the overall existence of personification metaphors, which construed this institution as a human being with the ability of performing both mental and physical tasks (e.g., Commission is aware, la Comisión desea reiterar que). Following these findings, our hypothesis is that personifications-with-metonymy are an important part of EU discourse, and the goal is to refine such observations maintaining similar methods (corpus-based extraction of verbal patterns with named entities) whilst adding a conceptual layer to the previous analysis, which was more syntactically oriented. To this end, we will focus on the personification of organisations (ORG). This is one of the most frequent types of named entity in our corpus (cf. Corpas Pastor and Sánchez Rodas 2022/In press), and it has been closely related to the different ways in which personifications-with-metonymy manifest in language use (Viimaranta and Mustajoki, 2020).

3 Methodology

The methods employed are like those in Corpas Pastor and Sánchez Rodas (2022/In press), that is, automatic extraction of patterns with named entities from an intermodal English-Spanish corpus managed with the Voice-text Integrated System for Interpreters, or VIP (Corpas Pastor, 2020). The corpus (PETIMOD 2.0) is composed of mediated and non-mediated citizens’ petitions and other working documents (summaries, minutes, opinions, etc.) from the European Parliament’s Committee on Petitions (PETI), plus transcribed meeting interventions of MEPs and invited speakers. More specifically, it comprises four subcorpora: Non-Translated English (N-T_EN), Translated Spanish (T_ES), Non-Interpreted Spanish (N-I_ES, and Interpreted English (I_EN) (Corpas Pastor and Sánchez Rodas, 2022/In press). This paper presents a monolingual analysis (N-T_EN + I_EN) because of space limitations. The tables below summarise the dimensions of PETIMOD 2.0 as counted by ReCor, in general and by language.

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>Documents</th>
<th>Types</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-T_EN</td>
<td>21</td>
<td>5330</td>
<td>52421</td>
</tr>
<tr>
<td>I_EN</td>
<td>81</td>
<td>2025</td>
<td>15709</td>
</tr>
<tr>
<td>N-I_ES</td>
<td>81</td>
<td>3072</td>
<td>18409</td>
</tr>
<tr>
<td>T_ES</td>
<td>21</td>
<td>6262</td>
<td>61377</td>
</tr>
<tr>
<td>TOTAL</td>
<td>204</td>
<td>16689</td>
<td>147916</td>
</tr>
</tbody>
</table>

*Figure 1. PETIMOD 2.0 size (per subcorpus)*

---

10 A software application for measuring corpus size and representativeness developed by the Lexytrad group (http://www.leyxtrad.es/en/resources/recor-3/).
Following Biel (2014), we use term-embedding collocations of the type N(term)+V to automatically extract possible personifications, given their conceptual relevance. More specifically, we employ the pattern ENT (=ORG) + V. Performance of the two models of VIP for automatic pattern extraction (SpaCy and DeepPavlov) is compared in terms of precision. To this end, precision errors include wrongly identified and/or classified NEs plus wrongly identified verbs.

<table>
<thead>
<tr>
<th>PETIMOD 2.0 (ORG + V)</th>
<th>SpaCy</th>
<th>DeepPavlov</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N-T_EN</td>
<td>I_EN</td>
</tr>
<tr>
<td>TOTAL ORG + V</td>
<td>105</td>
<td>32</td>
</tr>
<tr>
<td>ERRORS</td>
<td>31</td>
<td>3</td>
</tr>
<tr>
<td>RELEVANT ORG + V</td>
<td>74</td>
<td>29</td>
</tr>
<tr>
<td>PRECISION</td>
<td><strong>0.704</strong></td>
<td><strong>0.906</strong></td>
</tr>
</tbody>
</table>

Means show that DeepPavlov performs slightly better when it comes to pattern extraction in English (+0.06). Since internal differences are very small, we will use the DeepPavlov extraction for the detection of personifications.

To analyse the personifications in a qualitative and quantitative way, our last step is to follow Dorst et al. (2011) to manually separate the ORG + V collocations in which the verb has a basic human sense according to the Macmillan English Dictionary for Advanced Learners (Rundell and Fox, 2002). We retrieve the context of the collocations from the corpus concordancer and determine the humanness of the verb by looking at signal nouns and pronouns (Dorst et al., 2011: 177), signal verbs, and sentence examples in the Macmillan dictionary definitions. Signal verbs are also used to create a metonymical personification typology from the point of view of verbal usage.
4 Results and discussion

<table>
<thead>
<tr>
<th>PERSONIFICATION TYPE</th>
<th>N-T_EN</th>
<th>I_EN</th>
<th>EXAMPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTION</td>
<td>29</td>
<td>7</td>
<td>ENSREG prepared a template for the National Action Plan</td>
</tr>
<tr>
<td>BELIEF</td>
<td>3</td>
<td>3</td>
<td>The EESC hopes that this report will help EU countries</td>
</tr>
<tr>
<td>REASONING</td>
<td>9</td>
<td>2</td>
<td>the Assembly decided to harmonise its name in the various official languages</td>
</tr>
<tr>
<td>SPEECH</td>
<td>13</td>
<td>8</td>
<td>if the Supreme Court rejects the request for an annulment then obviously closing the petition</td>
</tr>
<tr>
<td><strong>TOTAL</strong> (IN RELATION TO RELEVANT ORG + V)</td>
<td><strong>54/80</strong> (67.5 %)</td>
<td><strong>20/28</strong> (71.4 %)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4. ORG Personification count for PETIMOD 2.0 N-T_EN and I_EN

Results show that the combination of automatic ORG+V pattern extraction with MIPVU is an effective method for detecting personifications in the English Eurolect, as more than a half of the patterns in both subcorpora (54 out of 80 in N-T_EN and 20 out of 28 in I_EN) are labelled as such). Although ORG+V patterns are generally less frequent in mediated discourse (probably because of simplification, as described in Corpas Pastor and Sánchez Rodas 2021/In press), personifications are relatively more common in this subcorpus, with a presence of 71.4 % (3.9 % more than in the non-translated texts). Interestingly, verb frequencies change between written and oral discourse, with a preference for verbs of speech and action over reasoning in the oral texts. This could indicate that in the oral discourse of the committee meetings there is a tendency for presenting institutions as people who say things (meta-referential or quoting function) and people who do things, rather than people who think. When combined, the four types of personifications deducted from the definitions in the Macmillan dictionary (institutions that act, believe, reason, and speak) seem to form the more comprising metaphor INSTITUTIONS ARE PEOPLE.

5 Conclusions

The method employed in this paper (automatic pattern extraction plus metaphor identification using dictionary definitions and context) for the analysis of humanised institutions in EU texts seems to open promising ways for the empirical study of personification in specialised language. It must be recalled, however, that this paper has focused almost exclusively on the metaphorical cross-domain mappings drawn by personification when institutions act, believe, reason, or say something. More studies are still needed to dig in the
interesting choices for one or another named entity in the corpus collocations (e.g., why the EESC can hope, but not the Commission). This can only be done from a constructional approach to language that bridges the gap between metonymy (the name) and metaphor (the verb), whose influence has already been envisaged in the dictionary definitions and the in-context analysis of the extracted personifications. In turn, a constructional description of the personifications found in our multilingual institutional corpus would delimitate a quite human, complex metaphorical mechanism in language, discovering possible differences and/or similarities between English and Spanish that could be accurately transferred to or evaluated in the performance of several NLP applications (machine translation and interpreting, for example).

Acknowledgements

The research reported in this paper has been funded by the former Spanish Ministry of Education and Professional Training, now Ministry of Universities (ref. FPU18/05803). The paper has also been carried out in the framework of the projects VIP (FFI2016-75831-P), TRIAGE (UMA18-FEDERJA-067), MI4ALL (CEI-RIS3), and VIP II (PID2020-112818GB-I00).

References


OCCAM: cross-lingual unlocking of non-digital texts

Laurens Meeus  
CrossLang, Ghent  
laurens.meeus@crosslang.com

Joachim Van den Bogaert  
CrossLang, Ghent  
joachim.vandenbogaert@crosslang.com

Arne Defauw  
CrossLang, Ghent  
arne.defauw@crosslang.com

Oan Stultjens  
CrossLang, Ghent  
oan.stultjens@crosslang.com

Sara Szoc  
CrossLang, Ghent  
sara.szoc@crosslang.com

Tom Vanallemeersch  
CrossLang, Ghent  
tom.vanallemeersch@crosslang.com

Frederic Everaert  
CrossLang, Ghent  
frederic.everaert@crosslang.com

Koen Van Winckel  
CrossLang, Ghent  
koen.vanwinckel@crosslang.com

Abstract

The OCCAM project (OCR, ClassificAtion & Machine Translation, 2019-2021), funded by the Connecting Europe Facility programme of the European Commission, integrates classification, optical character recognition and translation technologies into a single application to support multilingual access to scanned documents. The OCCAM application, currently including five languages (Dutch, French, German, Czech and English), is machine learning based, open-sourced, fully customisable and accessible through a user-friendly interface. The OCCAM workflow is highly relevant in various scenarios that show an urgent need for making non-digitised texts accessible in a multilingual context. We illustrate this with two use cases, i.e. the processing of image-based documents in business registers and the combination of digitising and translating historical texts in the area of digital humanities.

1 Introduction

Despite increasing digitisation efforts over the last decade, large amounts of valuable texts remain tucked away and inaccessible to many potential users and humanities researchers. Furthermore, making these documents accessible often depends on accurate optical character recognition (OCR) solutions, but also faces important language barriers, such as the original language of the document which may not be understood by the recipient or may hamper the further processing of such documents due to a lack of available multilingual tools. These challenges were addressed by the OCCAM project (OCR, ClassificAtion & Machine Translation), a collaboration between CrossLang (Belgium), the Brno University of Technology (Czech Republic) and the Ghent Centre of Digital Humanities (GhentCDH, Belgium), which ran from October 2019 to September 2021.

11 GhentCDH acted as subcontractor.
OCCAM integrates image classification, OCR and translation technologies in one application to support the automated translation of scanned documents and thus help increase multilingual access to documents that have not yet been digitised. When a document enters the workflow, it is first classified to identify the type of document. Then, the transcription is produced using one of the available OCR models. In a subsequent step, the transcription is fed to a translation module consisting of a translation memory (optional) and a machine translation (MT) system, that enriches the document with multilingual data. The components of the OCCAM application, made available as open source\(^\text{12}\), are based on machine learning (deep learning) and can be customised for different uses.

The project involved two use cases: (i) connecting the Business Registers Interconnection System (BRIS) of the European Commission (EC) with the EC’s eTranslation system using OCCAM to allow large volumes of image-based business documents that currently cannot be processed to become accessible in multiple languages, (ii) in the area of digital humanities (DH), the combination of digitising and translating historical texts leads to a significant growth and diversification of accessible corpora, thus helping to preserve cultural heritage and promoting the discovery of new knowledge and insights.

This article is structured as follows. Section 2 presents the OCCAM application and its graphical user interface (GUI). We elaborate on the application’s key components in Section 3, after which we describe the two use cases of the project in Section 4. We discuss challenges and future prospects in Section 5. Finally, Section 6 presents the main conclusions of the OCCAM project.

2 OCCAM workflow and user interface

As the OCCAM software is intended for practical applications in areas such as digital humanities and business administration, it has a GUI allowing the user to upload a (set of) scanned document(s), process them, and download the resulting transcriptions and translations. First-time users can take a short virtual tour guiding them through the application, or they can have a look at the video tutorial.\(^\text{13}\) A demo version of the application is also available for testing.\(^\text{14}\)

Processing a document in OCCAM is a five-step process:

1) Document Information: The user creates a new project by providing a title and, optionally, a summary of the document to be processed.

2) Upload Pages: In this view, the user uploads a document in PDF or image format. If the document consists of several files, they can all be uploaded at once. In the background, the type of uploaded document is detected and saved in the metadata.

Figure 1: GUI progress menu with the five workflow steps (views)

---

\(^{12}\) See the GitHub repositories on https://github.com/CrossLangNV.

\(^{13}\) A short tutorial of the OCCAM tool is available at: https://www.youtube.com/watch?v=SMt9re0g1es.

\(^{14}\) For a web application demo, please visit https://react.staging.occam.crosslang.com.
3) **OCR Model:** After successfully uploading the documents, the user is prompted to select an OCR engine among the different available models. The document type detected in step 2 is used to automatically suggest an optimal OCR model. When the user clicks on the model name, a brief description is given.

4) **Results:** This is the application’s main view. It allows the user to review uploaded pages, launch the OCR engine on selected documents, view the resulting transcriptions in various ways, and translate the transcriptions into the target language(s) of choice. The source language of the document is automatically detected.

5) **Publish:** Finally, the user can download the results in a zip file or may choose to publicly share the data through the Open Archives Initiative (OAI-PMH).

![Figure 2: Results view → Page View](image)

As shown in Figure 2, after a document is processed by the OCR module, text regions are recognised and overlaid on top of the image by blue boxes. When hovering over such a blue box, the corresponding transcription pops up. By navigating to the tab with the target language (English), the text region is overlaid with the translation of the transcribed text. The text transcriptions and translations can also be viewed as full texts by clicking the Text View tab.

### 3 OCCAM components

The OCCAM application has three key components: document classification, OCR and MT. In the next sections, we will elaborate on each of them to provide a better understanding of how they work and interact with each other.

#### 3.1 Document classification

When a document is uploaded to OCCAM, the tool provides an automatic suggestion about the optimal OCR setting for this document, which is based on metadata generated by a series of document classifiers.

---

15 https://www.openarchives.org/pmh
The automatic generation of metadata is illustrated in Figure 3 (to be read bottom-up). A first document classifier distinguishes machine-readable from non-machine-readable documents. Machine-readable files contain text that can be easily processed by the computer (e.g. a PDF file created from a Word document), while non-machine-readable files, usually originating from scanned documents or images, are restricted in their usability. In the example in question, the document is identified as a scanned document (i.e. as a non-machine-readable image). Next, the scanned page is classified as not belonging to the DH domain. Then, it is identified by a third classifier as being from the Belgian Official Gazette (BOG), rather than from the National Bank of Belgium (NBB). Finally, based on this information, the OCCAM application is now able to suggest an optimal OCR model to be used for this particular document. This is shown in Figure 4.

---

16 This classifier was specifically designed for the business registers use case in the framework of the OCCAM project to distinguish between documents from the BOG and annual accounts from NBB.
3.2 Optical Character Recognition

As shown in the previous Section, the OCCAM application provides access to a set of different OCR engines to make it possible to receive tailored results according to the type of document being uploaded.

The OCR component of OCCAM is based on PERO OCR, a novel machine-learning based OCR engine developed by the Brno University of Technology, to automatically transcribe several types of printed and handwritten documents. This OCR component has three main subcomponents, all of which can be customised for a specific document domain (Kišš et al., forthcoming).

- **Page layout analysis**: This subcomponent performs detection of text regions (paragraphs and text lines) in scanned documents. Models for printed, handwritten or combined documents are available (see Kodym and Hradiš, forthcoming).

- **Visual OCR models**: This subcomponent converts the detected text regions of a given document into text transcriptions. PERO-OCR provides models for the following document types: (i) prints and older typewritten documents, (ii) handwritten documents (available for Czech, English, German, Slovak, and old German), (iii) older prints in Fraktur, Schabach and other similar fonts, and (iv) Arabic printed and handwritten documents. Figure 5 provides an example of the processing of a handwritten document in the OCCAM application.

- **Text-only language models**: Automatic transcription of scanned documents may introduce various types of errors depending on the difficulty of the task at hand (type of script, language, etc.). To enhance the quality of the transcription, language models are used. The PERO-OCR system uses pretrained language models for modern English, modern Arabic and Czech.

![Figure 5: Handwritten document and its transcription in the OCCAM application](https://pero-ocr.fit.vutbr.cz)

For more information, please visit the PERO-OCR project website at https://pero-ocr.fit.vutbr.cz; the source code can be accessed via https://github.com/DCGM/pero-ocr.
3.3 Machine Translation

To make a document accessible in multiple languages, the OCCAM application provides the user with the possibility of translating the text transcriptions that were obtained by applying an OCR model.

In the framework of the OCCAM project, the application was connected to the CEF eTranslation system developed by the EC, specifically for translations involving French, English, German, Dutch and Czech. However, it is possible to extend the application to any other MT engine or language pair. As MT engines typically take sentences as input, the text in each paragraph is first segmented into sentences, which are then sent to the engine for translation. In addition, there is an option to upload and make use of a translation memory, i.e., a database storing sentences and their translations. The engine will first look for an exact match of a sentence within the memory. If it exists, the translation is retrieved and saved; all sentences that cannot be translated this way are translated by the MT engine. Figure 6 shows the line being hovered over in Figure 2, but now the English translation appears (as will be discussed in Section 5, the OCR quality can hamper the quality of the MT output).

![Figure 6: Results view → Page View with English translation](image)

4 Use cases

The OCCAM project involved two use cases: a workflow for retrieval of information from business registers and a workflow for DH researchers.

The first use case was investigated by reaching out to one of the EC’s Digital Service Infrastructures (DSIs), i.e., the Business Registers Interconnection System (BRIS). The platform provided by the DSI facilitates access to information on EU companies and ensures that all EU business registers can communicate with each other electronically. By querying this platform for information on specific companies, entrepreneurs can for instance more easily work on cross-border mergers. The DSI acknowledges the potential of integrating the OCCAM pipeline with their centralised platform, as large volumes of image-based business documents that currently cannot be processed would become accessible in multiple languages.

One way to integrate OCCAM with an existing platform for retrieving documents may consist in sending the retrieval results on-the-fly to the OCCAM pipeline. Besides providing

---

18 [https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation](https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation)
translated documents, the use of OCCAM may lead to value-added services such as mining retrieved and translated text names of entities (organisations, persons, etc.). It should also be noted that the OCCAM consortium ensured that its application guarantees the secure transfer of data, through the eDelivery software of the EC. This is especially important for business-related scenarios.

The second use case was investigated by reaching out to the DH community. In the last couple of years, manuscripts and other historical resources have become increasingly available as images or PDF files, for instance through digitisation of libraries. Although several OCR frameworks now exist that can convert these scanned documents into machine-readable text transcriptions allowing for full text search and alleviating the need for fully manual transcription, the demand for combining OCR with translation still remains pressing in the community. This combination facilitates a significant growth and diversification of accessible corpora, and thus the preservation of cultural heritage. It helps to link resources at various locations and in different languages, allowing cross-lingual searches, and provides leverage to discover new knowledge and insights.

In order to reach out to the DH community, the consortium contacted the Europeana DSI through the Ghent Centre of Digital Humanities. Europeana is the European flagship initiative for digital cultural heritage. The OCCAM software offers opportunities for the DSI to integrate OCR in a pipeline for translating metadata and full texts and provides search functionalities across languages. The DSI makes provision for the translation of the metadata in its collections through the recently started Europeana Translate project.19

5 Challenges and future avenues

While the two use cases in the project involved a restricted set of languages (Czech, Dutch, French, English and German) and consortium contacts with specific teams, they support the OCCAM project results as being sustainable in the long term. The technology developed can be applied widely in environments where multilingual access to registers is required, and researchers in universities, libraries, museums and archives need to process and access corpora in various languages and domains. The technology is publicly available as open source and its generic setup also allows for plugging in new OCR models and, if applicable, customised MT systems. An online tutorial for the DH use case is available.20

A major challenge in many natural language processing pipelines lies in the fact that the errors in the first system applied are propagated to the second system, leading to an accumulation of errors. This also pertains to the combination of OCR and MT. An MT system is designed for translating grammatical sentences produced by human writers. Given such sentences as input, it may produce some errors because of the inherent limitations of the MT system. When the input consists of sentences resulting from OCR, the input itself may also contain errors. This will lead to a combination of errors in the MT output. An example of the difference between MT output (from eTranslation) following incorrect OCR output and MT output following correct OCR output is shown in Figure 7. The problem of error propagation is especially prominent in small text regions identified by the OCR system, such as titles and table cells - due to the lack of context, such items are hard to recognise correctly for an OCR system and hard to translate for an MT system. Potential future strategies for mitigating OCR and MT errors and their propagation may consist of applying lexicon-based spell checking and grammatical correction to OCR output and devising procedures for character recognition which consider context beyond the text unit in question (sentence, title, etc.).

19 https://pro.europeana.eu/project/europeana-translate
20 https://www.youtube.com/watch?v=kL-kcdJSzOA


MT output: according to Bozner’s weekly markets of 31. December 1841.

Figure 7: Impact of OCR errors on MT output

Another major challenge is that older language variants not only require training specific OCR models but also specific MT models. Even given perfect OCR output but text that is not highly domain-specific, a standard MT model may not be able to translate some words or grammatical constructions that are obsolete in modern language. This is illustrated by Figure 8, which contains the words Baumann and Bedingnisse, more or less corresponding to the modern words Bauer (“farmer”) and Bedingungen (“conditions”). Providing texts in an older language variant may lead to unexpected behaviour and impact the readability of the MT output. Therefore, parallel data needs to be collected for the older language variant (texts written in that variant and their translation into the modern version of the target language), to train the MT model. Such a collection effort is notoriously difficult, and may in some cases stumble upon a total lack of appropriate (and accessible) data. Another possibility is to design systems that “translate” text from an older language variant into modern language, so tools developed for processing the latter, such as an MT engine, can be indirectly applied to text in the older variant; see for instance Tjong Kim Sang et al. (2017) for conversion of seventeenth-century Dutch into modern Dutch.

Figure 8: Text containing obsolete words, Baumann and Bedingnisse

Several potential extensions of the OCCAM software may be considered. Besides experimenting with the techniques explained above, it may be interesting to provide the user interface with functionalities for manual correction of the OCR output and the MT output and for retraining OCR and MT models. This would reduce the user’s dependency on a technical support team, and would also result in a larger user base and facilitate the demonstration of the software’s learning cycle to potential new users. More extensive use of the application may also lead to further requests for improvement of the software, beyond what was observed during the project.

6 Conclusions

While various OCR frameworks exist to support the conversion of scanned documents into machine-readable text, the variety of uses of scanned documents (retrieval, scientific text research, cross-border access) demonstrates a strong need for tailored OCR models and automatic translation of the transcribed content. Moreover, the practical steps required for transcribing and translating need to be straightforward to support a large user base.
In the OCCAM project, we integrated document classification, OCR, MT and translation memories into one application to unlock non-digital texts and make their content available in several languages. The components of the application we constructed are based on deep-learning, open-sourced and fully customisable. A user-friendly GUI allows the uploading of a set of scanned documents (PDF or image format), their processing, and the downloading of the resulting transcriptions and translations. The application first performs document classification to distinguish machine-readable from non-machine-readable content (printed or handwritten) and to identify different types of documents, in order to select an OCR model, which then performs detection of text regions and transcribes the text in these regions. The sentences in the transcribed text can be translated using MT. The user can view the transcriptions and translations either in separate tabs, or as an overlay on top of the image, by hovering over parts of the image. If a translation memory is provided and a sentence is found in it, its translation is retrieved instead of sending it to the MT system.

In the project, two use cases were elaborated, i.e., the processing of image-based business documents, and the combination of digitising and translating historical texts in the area of DH. The communication with two of the EC’s DSIs, BRIS and Europeana, indicated that the OCCAM application is potentially useful for their environment. The former offers a centralised platform facilitating access to information on EU companies, while the latter needs to provide the users of its collections with translations of metadata and full texts as well as cross-lingual search functionalities. While the two use cases involved a restricted set of languages and a specific MT system (CEF eTranslation of the EC), they confirm the OCCAM project results as being sustainable in the long term. The technology developed can be widely applied in environments involving multilingual access to registers and to DH corpora in various languages and domains, and can be adapted to support various MT systems and language pairs.

Major challenges, providing future opportunities for extending the OCCAM application, consist in tackling error propagation from the OCR to the MT system, and applying MT systems to older language variants.

Acknowledgements

OCCAM is funded by the EC’s CEF (Connecting Europe Facility) Telecom programme (2018-EU-IA-0052).

References

