Does training in post-editing affect creativity?

Guerberof-Arenas, A.; Valdez, S.; Dorst, A.G.

Does training in post-editing affect creativity?

Ana Guerberof-Arenas, University of Groningen
Susana Valdez, Leiden University
Aletta G. Dorst, Leiden University

ABSTRACT

This article presents the results of an experiment with eleven students from two universities that translated and post-edited three literary texts distributed on the first and last days of their translation technology modules. The source texts were marked with units of creative potential to assess creativity in the target texts (before and after training). The texts were subsequently reviewed by an independent professional literary translator and translation trainer. The results show that there is no quantitative evidence to conclude that the training significantly affects students' creativity. However, after the training, a change is observed both in the quantitative data and in the reflective essays, i.e. the students are more willing to try creative shifts and they feel more confident to tackle machine translation (MT) issues, while also showing a higher number of errors. Further, we observe that students have a higher degree of creativity in human translation (HT), but significantly fewer errors in post-editing (PE) overall, especially at the start of the training, than in HT.

KEYWORDS

Translation training, machine translation post-editing, creativity, creative shifts, acceptability.

1. Introduction

Since the integration of personal computers and computer-aided translation (CAT) tools into the translation profession, scholars have reflected on how technology should be included in translation training (Bowker 2002; O'Brien 2002; Pym 2014; Kiraly, Massey and Hofmann 2018). Although to varying degrees, mainly depending on cultural preferences and translation schools, these tools are now included in most curricula and are generally considered essential for up-to-date and rounded translation training. The DigiLit (2023) project, for example, looks to "determine how the advantages of AI-related technology can be leveraged in language teaching and academic writing while minimizing potential issues such as ineffective communication, miscommunication, and misuse" (e.g. plagiarism).

In the last fifteen years, the commercialisation of MT and its adoption by the localisation industry have created the need to also include training modules dealing with MT technology – in different paradigms such as ruled-based, statistical or, most recently, neural MT (NMT) – as well as post-editing (PE) practice (Doherty and Moorkens 2013; Doherty and Kenny 2014; Guerberof Arenas and Moorkens 2019). Knowing how to use MT effectively is recognized as an essential skill for future translators (Rothwell 2019; EMT Board and Competence Task-Force 2022), and students more generally (Bowker 2020; Dorst, Valdez and Bouman 2022; Loock, Léchauguette and Holt 2022).

https://doi.org/10.26034/cm.jostrans.2024.4712

The recent developments of NMT technology and the increased quality of its output require translation trainers to reflect on the future skills for the next generation of translators. Regardless of assurances that machines are not going to replace translators any time soon (Way 2019; Nord 2023), there is a perception in the classrooms that the "rise of the machine" is inevitable, resulting in the fear that training might become irrelevant once full automation replaces the translator. Further, with the recent developments of Large Language Models (LLMs) paired with conversational agents such as ChatGPT or BingChat, the media (Heaven 2023) and part of the AI community appear to classify translation as a replaceable activity (Eloundou *et al.* 2023).

However, if machines show such effectiveness for standard and simple texts that need, in turn, less human intervention, then translators need to focus on more unusual and complex texts (King 2019). In other words, translators need to become more creative to show their "added value" or "advantage" over the machine. Recent research shows that PE constrains creativity in the translation of literary texts (Guerberof-Arenas and Toral 2022), resulting in poorer translations. But does training in the use of MT and PE help to increase or decrease students' creativity? Will the current training in technology result in more problem-aware translators who can do PE while still maintaining their creativity?

In this article, we look at the results of a joint experiment at two universities, University of Groningen and Leiden University. Eleven students translated three texts distributed on the first and last days of the modules. The source texts were marked with units of creative potential to assess creativity in the target texts (before and after training) following the methodology developed by the CREAMT (2022) project. The texts were subsequently reviewed by a professional literary translator and translation trainer. We present here the results and our reflections from this experiment that seeks to answer the question: What is the effect of MT and PE training on students' creativity?

2. Creativity, PE and competence in translator training

In this section, a summary of the literature regarding existing training with a focus on creativity and PE will serve to contextualise the current study. Due to space limitations, it is not possible to cover all the research done on translation or MTPE training in recent years.

2.1. Creativity in translation

Creativity is seldom studied in the intersection with technology or included as a requirement of the translator training curriculum, even if it is often discussed during translation classes. This focus on creativity is becoming increasingly necessary in view of technological advances in MT but also the increasing fluency and accuracy of LLMs such as GTP-3 or, more recently, ChatGPT (that avails of reinforcement learning from human feedback).

The research presented here is included within the framework of the CREAMT project and, therefore, uses the same instruments. This project aims to identify and quantify creativity in relation to technology and determine the impact of creativity on translators and readers. Seminal research by Paul Kussmaul (1991; 1995) used the concepts of scenes and frames (Fillmore 1977) to explain how translators transpose these frames and scenes in the target language by using *shifts*. Later, Gerrit Bayer-Hohenwarter coined the term *creative shifts* and *reproduction* (Bayer-Hohenwarter 2011) and contributed to the measurability of creativity in translation, i.e. how translators resolve a unit of creative potential (a translation problem).

The CREAMT project went further by creating a complete taxonomy of units of creative potential and of creative shifts, reproductions and omissions that is used here and explained in detail in Section 3. This research also found that when professional translators work without MT, they are more creative than when post-editing, meaning that they tend to use more creative shifts and make fewer errors (Guerberof-Arenas and Toral 2022). To our knowledge, there are no experiments that test students' creativity in their intersection with technology, and, hence, this work aims to fill this gap.

2.2. PE training

CAT tools and MT have, at least in part, been the industry's solution to what Dunne calls the "productivity imperative" (Dunne 2012, 155). Translators and, by extension, language service providers are expected to translate and localise larger volumes of text in shorter turnaround times at a lower price. Eventually, this need has pervaded translator training in Higher Education as demonstrated by the number of published articles on this topic (O'Brien 2002; Doherty and Kenny 2014; Kenny and Doherty 2014; Mellinger 2017; Guerberof Arenas and Moorkens 2019; Nitzke, Tardel and Hansen-Schirra 2019).

In one of the first articles on MTPE training, O'Brien (2002) proposes an outline for a course module in PE. O'Brien (2002, 101) argues that, in PE, students need to understand early on that there are diverse levels, depending on clients' needs and learn to "use as much of the raw MT output as possible" to take advantage of productivity gains. To tackle this issue, O'Brien proposed a syllabus that includes a theoretical component and a practical component, including PE practice, terminology management, coding, controlled language, corpus analysis, and programming.

Kenny and Doherty (2014) and Doherty and Kenny (2014) argue that it is essential for students to understand the value of keeping up to date with industry standards. They propose a statistical machine translation (SMT) syllabus that includes, among other topics, basic knowledge of SMT, MT evaluation, pre- and post-processing, and human and professional issues in MT such as ethics and payment. Despite the fact that the MT paradigms have changed, these topics are still relevant as part of the training programme adapted to the ever evolving technical landscape.

Instead of proposing stand-alone modules or courses, Mellinger (2017) argues for cross-module or cross-curricular integration of MT to foster the linguistic and technological competences that graduates will need to succeed in the translation industry. At the base is the argument that if MT is included across various modules, translator trainers can address aspects often neglected at different stages and within the context of different topics (translation practice, revision or writing courses). He goes on to propose four topics to be implemented curriculum-wide: terminology management, controlled authoring, PE, and engine tuning.

Guerberof-Arenas and Moorkens (2019) describe an MT and PE course, as well as an MT project management module. There is a considerable focus on giving students "a realistic view of the task" and on "expand[ing] on different concepts of quality in localisation, not as a universal value but as a 'granular' concept depending on customer requirements" (2019, 222–223), through project-based assignments. They propose a PE module that covers basic definitions of PE, quality, types of PE, guidelines for light and full PE, common MT errors, PE effort and productivity, and PE and pricing.

2.3. PE competence in the EMT Competence Framework perspective

The ability to use MT has been part of the European Master's in Translation (EMT) Competence Framework since its inception in 2009. This signals that the use of MT, and by extension of PE, was recognized early on by one of the leading standards for translator training in the EU, if not beyond. "Knowing the possibilities and limits of MT" was already contemplated as part of the technological competence for professional translators and experts in multilingual and multimedia communication (EMT expert group 2009, 7).

In the 2017 EMT Competence Framework, this knowledge was further expanded; from being restricted to a technological competence, the ability to post-edit MT output was recognized for the first time as an "integral part of professional translation competence" (EMT Expert Group 2017, 17).

In its recently revised version, the 2022 EMT Competence Framework (adopted for 2023-2028) expands yet again the scope and breadth of this competence. Under translation competences, the "heart" of the competences represented in the framework, MT is acknowledged as a "growing" but also constituting part of translation workflows. MT literacy is also referred to for the first time, due to the increasing research outputs on the topic and the associated societal impact of MT usage (e.g. Bowker and Ciro 2019; Delorme Benites *et al.* 2021; Kenny 2022; Krüger and Hackenbuchner 2022).

3. Methodology: Measuring creativity in translators' training

As we mentioned in Section 1, we wanted to know how students' creativity was affected, positively or negatively, after undergoing training on MT and PE as part of a Translation Technology module at their Master programme.

In the following sections, the design and instruments used to measure creativity before and after training are described.

3.1 Participants

University of Groningen students: Four students participated in this experiment. They have a Bachelor' s Degree in languages and a minor degree in Translation, and are in the first semester of their Master's in Linguistics, track Translating in Europe. As part of their course, they have modules on Translation Skills, Translating for the European Union, Translation Technology, and Comparative Grammar.

Leiden University students: Seven participants completed both assignments and were included for analysis. They are all master's students taking an obligatory 5-ECTS translation technology course (The Translator's Tools) as part of the 1-year Master's in Linguistics, track Translation. They have a Bachelor's Degree in languages and a minor degree in Translation. The EMT master's in Translation includes 3 obligatory 5-ECTS courses on translation theory, translation technology and advanced professional translation, as well as elective specialisation courses in Literary Translation, Legal Translation, Medical Translation or Multimodal Translation and Subtitling.

3.2 Overview of experimental design

We follow a pre-training and post-training design for this experiment. The same assignments were administered in the first and last weeks of their translation technology course[1]. The students had to translate and post-edit three texts taken from Hemingway's collection of short stories *In Our Time*[2] into Dutch. Three particular stories were chosen because of their length; they had a self-contained story and provided sufficient translation challenges for students. More importantly, the texts were available copyright free at Project Gutenberg. Chapter 2 (186 words) is a recount of a bull-fight, Chapter 6 (131 words) is a brief scene from a shooting of Greek ministers during the war, and Chapter 18 (151 words) is a scene that analyses the shooting of the ministers between the King and Queen of Greece and an unknown interlocutor.

Each assignment consisted of one short excerpt to be translated manually in MS Word, and two short excerpts that had been pre-translated using the free online MT tool DeepL[3] and needed to be post-edited. This MT system was chosen because it is one of the preferred public engines in the Netherlands and, therefore, students would be exposed to it in their professional activities.

As mentioned before, two groups were created in each university, so that the text difficulty could be balanced among the students and modality in the pre- and post-training phases. In the first class (pre-training), students in Group A first translated Chapter 2, and then post-edited Chapters 6 and

18, while students in Group B first translated Chapter 6, and then post-edited Chapters 2 and 18. In the last class (post-training), the groups were switched: Chapters 2 and 6 were done in the reverse modality. Table 1 illustrates the workflow:

Code	Pre-training	Group	Post-training	Group2	University
T1	**Saskia**	**A**	**Emma**	**A**	**RUG**
T2	Anna	A	Trees	B	RUG
T3	Gerrit	A	Robin	B	RUG
T4	Anika	B	Julia	A	RUG
T5	**Kaya**	**B**	**Layla**	**B**	**LU**
T6	Beatrice	A	Ellis	B	LU
T7	**Teddy**	**B**	**Rachel**	**B**	**LU**
T8	Jorge	A	Pedro	B	LU
T9	Faruk	A	Emre	B	LU
T10	Enrique	A	Sam	B	LU
T11	Kysia	A	Tina	B	LU

Table 1: Pre-training and Post-training assignments per translator with their code names

Unfortunately, 3 of the 11 eleven students that took part in the experiment (marked in bold in Table 1) did not follow the instructions and repeated the same modalities for the same chapters. However, we decided to maintain this data for our statistical analysis as the modalities were analysed separately. Further to the translation and PE assignment, students were asked to write reflective essays guided by questions in the first class (pre-training) and last class (post-training).

Each assignment from each translator in both cohorts was then coded and assigned a fictitious name (one in the pre-training and another one in the post-training as seen in Table 1), so that, in the end, 22 texts were reviewed by an independent professional literary translator (she is also a translation trainer) for both novelty and acceptability. Although the ideal would have been to use annotations from more than one independent reviewer to minimise the subjectivity of the evaluations, this was not possible due to budget limitations.

3.3 Data analysis: creativity

In order to analyse the data we looked at novelty and acceptability. The analysis of these two concepts is described in the following sections.

3.3.1 Novelty: units of creative potential and creative shifts

Creativity was defined as a combination novelty (i.e., new, original) and of acceptability (i.e., something of value, fit for purpose) as defined in previous

research on creativity in literary translation (Guerberof-Arenas and Toral 2022).

To assess novelty (creative shifts) to the given problems (units of creative potential, UCP), we first annotated these units in the ST, i.e. these are units that are expected to require translators to use problem-solving skills, as opposed to those that are regarded as routine units (Fontanet 2005; Bayer-Hohenwarter 2011).

To this end, the three researchers[4] annotated the texts following a UCP classification list, although they were free to mark others, too: A) metaphors and original images, B) comparisons, C) idiomatic phrases, D) wordplay and puns, E) onomatopoeias, F) colloquial language (cursing, slang, for example), G) phrasal verbs, H) cultural and historical references, I) neologisms, J) lexical variety (number of adjectives before the noun or use of adverbs, for example), K) expressions specific to linguistic variant (for example, American English or British English), L) unusual punctuation, M) rhyme and metrics, N) proper names, and O) treatment (formal, informal). All researchers have English as their second language, one has Dutch as their native language, and two have a B2 competency in Dutch according to the Common European Framework of Reference for Languages[5].

After annotating the text, the classifications were discussed and the researchers agreed on a final list containing 50 UCPs in the 3 selected chapters: Colloquial Language (4), Comparisons (1), Cultural and historical references (4), Expressions from linguistic variants (4), Idiomatic expressions (8), Lexical variety (6), Onomatopoeias (2), Phrasal verbs & syntactical expressions (11), Proper Names (3), Rhyme and Metric (2), Unusual Punctuation (5)[6].

After the UCPs were annotated, the target texts were sent to the reviewer who classified the UCPs in the target texts (TTs) according to this classification:

1) Reproduction: All translation solutions that reproduce the UCP with the same idea or image, even if they are acceptable. They can then be classified into Retention, Specification, Direct Translation or Official Translation.
2) Omissions: When a term or expression from the UCP is omitted in the TT. An omission can be subclassified as Creative or a Shortcut solution.
3) Errors: If the translation is not acceptable (contains too many errors to be classified), then it can be marked as Not Applicable (NA).
4) Creative shifts (CS): All translations that deviate from the ST in any of the following ways:

   —"Abstraction" refers to instances when translators use more abstract TT solutions. An abstraction could be subclassified into Superordinate Term or Paraphrase.

   —"Concretisation" refers to instances when the TT evokes a more explicit, more detailed and more precise idea or image. A Concretisation could be classified into Addition or Completion.

   —"Modification" refers to instances when translators use a different solution in the TT (e.g. express a different metaphor without the

image becoming more abstract or concrete). A Modification could be subclassified into Cultural, Situational or Historical.

3.3.2 Acceptability

To annotate the errors in the translated texts, the reviewer used the harmonised DQF-MQM Framework (Panić 2019). The errors were classified according to the following categories: Accuracy, Fluency, Terminology, Style, Design, Locale Convention, Verity and Other. The reviewer also had to annotate the severity of each error: Neutral (0 points for repeated errors or preferences), Minor (1 point), Major (5 points) and Critical (15 points). Kudos was used for exceptionally good translation solutions. The reviewer was sent instructions on how to perform each task according to the error descriptions already present in the DQF-MQM error taxonomy.

3.4 Reviewer

The reviewer has more than 15 years' experience as a translator and between 5 and 10 years as a professional translator trainer. Her native language is Dutch, and she works from Italian, English, French and German (in that order). She has translated approximately 40 novels on her own or as part of a team.
The reviewer is experienced with the methodology presented here as she participated in another experiment within the CREAMT project.
With the aim of eliciting her perceptions regarding the eleven students' performance before and after training, the reviewer was also asked to rate each student's performance using a 7-point Likert scale from 'Extremely bad (1)' to 'Extremely good (7)' followed by a justification of rating in an on-line questionnaire. The reason why we chose a 7-point Likert scale was to obtain more detailed feedback between the pre- and post-training activities since we suspected that the same participant would not have dramatically different results in a short period of time.
Each student's TT was assigned a fictitious name (see Table 1), and the reviewer was then asked to compare the output of two students at a time when, in fact, she was comparing the pre-training and post-training translations of the same student.

4. Results

In this section, the quantitative results are presented, including an analysis of the translated content by the reviewer, followed by an analysis of the reflective essays by the students.

4.1 Quantitative results

Since creativity is a combination of novelty and acceptability, the results are divided into these two categories[7].

4.1.1 Creative shifts

There are 50 UCPs identified by the annotators for the three texts per translator (11). Because of the experimental design, this meant that there are in total 1100 UCPs, divided as follows: 388 in HT and 712 in PE (35.3% vs 64.7% of the total). However, the number of UCPs between pre-training and post-training is equal: 550 in each phase.

In order to analyse the classification by the reviewer (the number of creative shifts, omissions, reproductions or unclassifiable UCPs) per modality and phase, and since the number of UCPs is not equal per modality, frequency tables are used. Table 2 shows the relation between Modality and Classification.

Classification	**HT**	**PE**	**Total**
CS	115 (29.6%)	160 (22.5%)	275
Reproduction	213 (54.9%)	507 (71.2%)	720
Omission	51 (13.1%)	35 (4.9%)	86
N/A	9 (2.3%)	10 (1.4)	19
Total	388 (100%)	712 (100%)	1100

Table 2: Frequencies according to classification and modality

The results show that there is a higher frequency of CSs and a higher number of Omissions in HT than in PE. The Chi-Square test (X^2 (3)= 38.34, $p < .05$) shows that these two variables are indeed dependent. Cramer's V indicated a weak association between the variables (V = 0.187). In the case of Omissions, these could be sub-classified as Shortcuts or Creative. In this case, from the 86 total Omissions, 61 are Shortcuts and 25 Creative ones (6 with errors). If we look at the modality, HT has a higher number of Omissions/Shortcuts (33) but also of Omissions/Creative (18) while PE has 28 and 7 respectively.

Table 3 shows the relation between Phase and Classification depending on the period of the technical training.

Classification	**Pre-training**	**Post-training**	**Total**
CS	130 (23.6%)	145 (26.4%)	275
Reproduction	374 (68.0%)	346 (62.9%)	720
Omission	36 (6.5%)	50 (9.1%)	86
N/A	10 (1.8%)	9 (1.6 %)	19
Total	550 (100%)	550 (100%)	1100

Table 3: Frequencies according to classification and phase

The results show that there is a higher frequency of CSs in the Post-training than in the Pre-training phase, and a higher number of Omissions. However, the Chi-Square test shows no significant values between these two variables. There is a higher number of Omissions in the Post-training of which 32 are Omissions/Shortcuts and 18 are Omissions/Creative (4 errors), while in the Pre-training, out of the 38 Omissions, 29 are Shortcuts and 7 are Creative (2 errors).

Therefore, looking at these two variables (Modality and Phase) independently in relation to the classification of the UCPs, the quantitative data shows that Modality, rather than Phase, is a factor affecting the classification of CSs, and that HT has a higher number of CSs.

To explore how both variables interact with each other and with the random factors (translators) a Generalised Linear Mixed Effects Model for Binomial Data[8] was fitted taking the dependent variable Creative_Shift, its interaction with the fixed variables Phase and Modality and with the random factor Translator. We find a main effect of Modality, with CSs less likely to occur in the PE condition than in the HT condition (β = -0.37, SE = 0.14, z(1100) = -2.59, p<.001). There are no significant effects of Phase, although there are fewer Omissions in the Pre-training condition[9].

Further, taking Omissions as the dependent variable, Phase and Modality as the fixed factors and Translator as the random factor we see a main effect of Modality, with Omissions less likely to occur in the PE condition than in the HT condition (β = -1.09, SE = 0.23, z(1100) = -4.72, p<0.00). There are no significant effects of Phase, although there are fewer Omissions in the Pre-training condition.

This data seems to support previous research where HT shows a higher number of CSs when compared to PE (Guerberof-Arenas and Toral 2022). It also shows a higher number of CSs after training but this is not significantly different between these two phases, so perhaps sufficient time has not elapsed to measure this variable. However, we see that there are also more Omissions in the Post-training phase, and the majority of these are Shortcuts.

When analysing the UCPs classified by the reviewer, it was observed that the Comments section also described errors. These UCPs were then sub-classified as Without errors and With Errors. Table 4 shows the relation between Classification and Errors.

UCP Classification	Without errors	With Errors	Total
CS	45 (6.4%)	230 (58.5%)	275 (25 %)
Reproduction	612 (86.6%)	108 (27.5 %)	720 (65.5%)
Omission	50 (7.1 %)	36 (9.2%)	86 (7.8%)
Not classifiable	0 (0%)	19 (4.8%)	19 (1.7%)
Total	707 (100%)	393 (100%)	1100 (100%)

Table 4: Frequencies according to Classification and Errors

The Chi-Square test (X^2 (3)= 445.18, p < .00) shows that these variables are indeed dependent. Cramer's V indicated a strong association between the variables (V = 0.636). This is different from previous research with professional translators, where the number of CSs did not lead to a higher number of errors, and this is to some extent logical since students have less experience translating and they are experimenting with new knowledge. To see how the modality and the phase affected the number of errors overall, the errors are analysed in more detail in the following section.

4.1.2 Acceptability

The reviewer classified all errors for the 11 translators into single DQF-MQM datasheets. These forms were aggregated to one single database that contains 792 responses (i.e. N = 36 sentences x 11 translators x 2 phases). As before, there are more responses in PE (578) than in HT (214). Figure 1 shows the number of error points (these include errors and their severity) according to the phase and the modality.

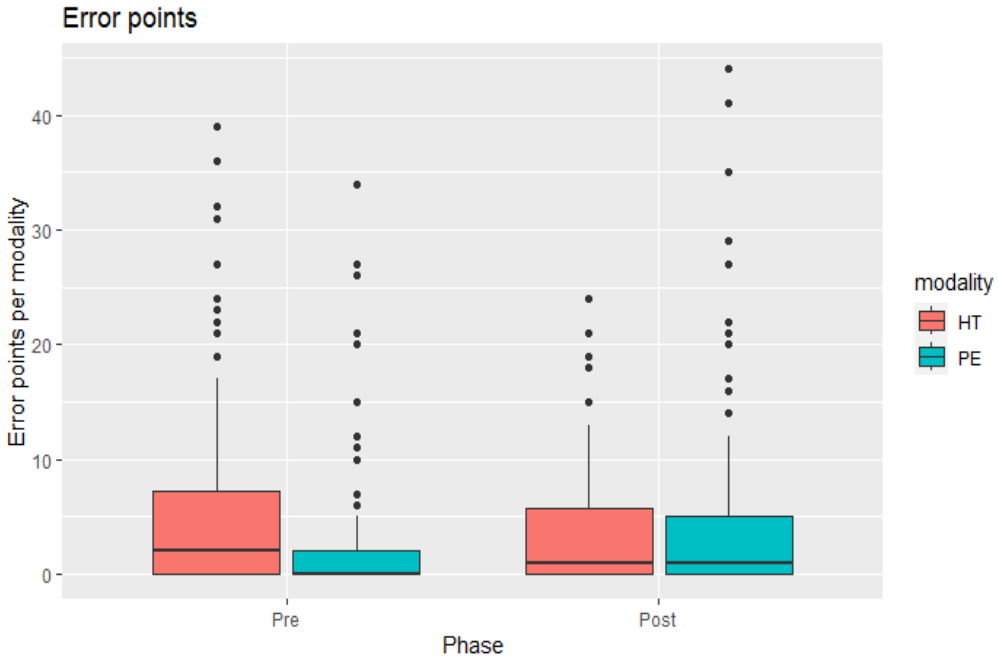

Figure 1: Error points per modality and phase

The results show that in the Pre-training phase (on the first day of the module) there are fewer error points in PE than HT. Previous research has shown that MT had a levelling effect on more novel translators (Guerberof Arenas 2014); in this case, it appears that PE helps novel *students* to have fewer errors. However, the errors increase in the Post-training phase and this is particularly noticeable in the PE modality, in the HT modality translators make slightly fewer errors than in the Pre-training phase. This could be explained by the fact that students were more careful in the first task than in the last task (it was the last day of the course and the assignment was not to be graded so this could have been done in a rush). However, their progression appears to be better in HT than in PE. Nevertheless there are fewer errors in PE than in HT in both phases.

To analyse this data more in detail and since the data is not normally distributed and this is an intra-subject design, we fitted a general linear mixed-effects model. The data that we are analysing was over-dispersed,[10] therefore we used a negative binomial family. The dependent variable Error_point was analysed in its interaction with the fixed variables Phase, Modality and Wordcount, and with the random factors Translator and ID_sentence. We find a main effect of Modality with Error_point is less likely to occur in the PE condition than in the HT condition ($\beta = -0.72$, $SE = 0.11$, $z(792) = -6.58$, $p<0.00$), and also in the sentences with a lower word count ($\beta = 0.09$, $SE = 0.02$, $z(792) = 4.35$, $p<0.00$). We see no significant effects of Phase, although there are fewer error points in the Pre-training condition. This is not in line with previous research with professional translators (Guerberof-Arenas and Toral 2022) where they show fewer errors in HT than in PE, but it is in line with previous research with students and/or non-professional literary translators (Stasimioti and Sosoni 2022).

4.1.3 Reviewer's rating of participants

Table 5 shows the ratings given by the reviewer to the students in the post-review online questionnaire. The reviewer was asked to rate the output on a 7-point Likert-type scale that ranged from Extremely bad (1) to Extremely good (7) in a blind paired student questionnaire distributed using Qualtrics[11].

Participant	Pre-rating	Post-rating
T1	5	6
T2	5	5
T3	3	2
T4	5	5
T5	4	5
T6	4	3
T7	2	2
T8	3	3
T9	6	6
T10	3	2
T11	6	5

Table 5: Students' ratings

Table 5 shows that 5 students maintain their quality (T2, T4, T7, T8, T9), 4 declined (T3, T6, T10, T11) and 2 improved (T1 and T5) but this difference fluctuates by one point only. Therefore, there are no dramatic changes in the students' performance as perceived by the reviewer.

In the reviewer's answers to the open question, the common denominator across all translation tasks, independently of whether they were conducted pre- or post-training, was mistranslations. These mistranslations were often found in Chapters 2 and 18. In the case of two of the participants, the mistranslations were derived from not considering the context. Another common observation was regarding unidiomatic translations. Four out of the eleven students produced unidiomatic translations and only two of the post-training translations were considered idiomatic or "fairly" idiomatic. Finally, the other recurrent observation was regarding overcorrections. In the case of two translations, the reviewer pointed out that their translations included "unnecessary rewriting" that, at least in one case, did not "influence style too much".

4.2 Reflective essays

This section covers the reflective essays written by the students both before and after training. A more detailed analysis of Chapter 18, which was post-edited in both assignments, can be found in the online repository for the CREAMT project[12]. After completing the assignment, the students were

asked to write a short reflective essay (max. 300 words) in response to a number of questions described herein.

4.2.1 Student reflections before training

In the pre-training phase (after completing the first assignment), the students were asked to reflect on the following questions: 1) How was the quality of the existing output by DeepL? 2) What was the nature of the changes made? 3) What resources did you use for the translation and post-editing? 4) In your opinion, which of three translations was the fastest to produce? 5) Are you satisfied with the final quality? 6) Which of the three translations are you the happiest with? 7) Did you like using MT as part of the translation process? and 8) How was translating different from post-editing?

In relation to the quality of the DeepL output provided, the student reflections show that while all of the participants were working with the same output, their reactions range from very positive to quite negative. While they all post-edited the same output, some found the quality "surprisingly good" (T4) or "better than I expected" (T11), while others considered it "quite lacking" (T3) or even "not publishable at all" (T2). Others referred to it as "mediocre" (T1), "quite alright" (T5) or "generally good" (T7). From the reflections in the pre-training phase, we can tentatively conclude that most of the students were expecting the output to be of lower quality and to contain more errors, given that half of the students (T4, T6, T8, T9, T11) explicitly referred to either being surprised by the quality of the output or the quality exceeding their expectations. This shows that exposure to MT output and determining just how many problems it contains is already a valuable learning experience for the students, which will help them determine whether MT is suitable for their text and gain a sense of whether they can trust the output.

In terms of the problem areas and error types they identified during their first attempt, one main problem with MT was that it is often "too literal", either in terms of lexical choice or in terms of word order. As one student puts it: "At times, the machine translation copied the sentence structure of the original text quite literally, sacrificing flow for accuracy." (T3). Another points out that she made mostly lexical changes, because "words were translated too literal [sic]" (T4). As far as the nature of the changes required during PE is concerned, 8 participants referred to the output needing lexical changes, 4 mentioned syntactic changes, 4 mentioned grammatical changes, 2 referred to naturalness, 1 referred to terminology, 1 referred to spelling, and 1 to retaining ST features (in this case, style).

Both the error labels used in the reflections and the examples provided do, however, show that not all students have the same knowledge of linguistics or meta-language to correctly identify and label errors and changes. For example, T11 argues that she had to "make some syntactical edits, such as

changing *revolutionaire* to *revolutionair*", and T8 explains that "Most were syntactical changes so congruency of adjectives and the nouns. Think of '*hele goede man*' instead of '*heel goede man*'".

One area of MTPE training that modules may want to focus on is therefore to provide students with more training in linguistics and error analysis. The question remains whether students can only become competent post-editors if they can correctly identify and label errors, or whether competence in PE can also be trained and improved without the meta-language and explicit linguistic knowledge. Our experience in MTPE training has been that students often make unnecessary changes and fail to make necessary changes, sometimes even adding errors. Some students admitted during class discussion that they often make simple changes because they do not see any errors and worry that the lecturer will think they did not do anything.

Another under-researched area in MTPE training is how students use resources. While studies on MT literacy consistently show that students use MT as if it is a dictionary (both monolingual and bilingual) and thesaurus (Dorst, Valdez and Bouman 2022), it remains unclear what kind of information students actually use from the dictionaries and other resources. In the reflections, students mentioned using online dictionaries, including Van Dale (4 times), Linguee (3) and Reverso (1), but also general websites like Wikipedia or Onze Taal (7), as well as MT websites such as Google Translate, DeepL and Reverso (6). Yet some of their explanations show that they are not using these resources to support their decision but rather they appear to be based exclusively on their own proficiency and intuition.

A clear majority of the students (8 out of 11) liked the experience of using MT and doing PE. One obvious reason was of course that PE is much faster than translation from scratch. Seven students mentioned explicitly that PE was faster than translating, and three noted the fact that PE saves time and allows you to get larger projects done quickly. Some students mentioned that it is faster to check an existing translation than to come up with a new one, and PE also requires less searching. Some also liked that the output gives them a rough basis to start from and offers alternative solutions: "it comes up with translations that I usually don't think of" (T11). Some students indicated that they felt the output was sometimes better than their own work, and sometimes they felt insecure about the correctness of the output (mentioned 4 times) or worried that they might be overlooking mistakes: "it is easier to overlook certain mistakes (such as strange colloquialisms) when the translated text is already before you" (T6). Another drawback they noticed was that they do not know whether to change the output if it uses words or constructions that do not appear to be incorrect but are not something they would use themselves (mentioned twice).

In the end, most students were happy with the results of their first attempt at MTPE and six were satisfied with the final quality. However, some did note explicitly that they preferred HT to PE because they felt more proud of the results and more certain of the translation's quality: "I'm satisfied about the final quality of the translations but I am happiest with the first one because I feel the proudest of that one, having made it from scratch" (T2). One student does note that "[e]ven after the post-edited machine translations have been edited, they still feel more rigid and unnatural than a translation that has been produced organically." (T3). Overall, five of the participants stated they preferred HT over MT; only one preferred MT over HT, because doing PE gave him more time to focus on style and content instead of grammar and lexis.

4.2.2 Student reflections after training

In the post-training phase (after completing the last assignment), the students were asked to reflect on the following questions: 1) During the task, have you noticed any changes in the way you post-edit the MT outputs or you translated the text, compared to when you did these tasks in the first session of this course?; 2) Are you satisfied with the final quality?; 3) Which of the three translations are you the happiest with? and 4) Did you like using MT as part of the translation process?

Reflecting on whether they did things differently, two students (T4, T9) remarked that they did not feel they did anything differently in the last week as compared to the first week. Three (T1, T2, T5) said they made fewer changes after training, while two (T6, T11) said they made the same number of changes, and one (T3) said they made more changes but this took less time. One student noted that even though they "edited approximately as much text as the last time" and "the changes I made were mostly the same", they were "much more sure about my choices" and "this time I was aware of what kinds of mistakes to look out for" (T11). This increased awareness and confidence was in fact the most noticeable change between the pre and the post-training phases. This is in line with previous research that looks into meta-cognition through reflective essays from students and finds greater awareness of the task at the global (content knowledge) and local (terminology and register) level over a time span of eight-week training (Mellinger 2019, 618).

Two students pointed out that the training had made them more aware of the task itself and their role as translator/post-editor: "I have not noticed a lot of difference in the way I translate of post-edit in the case of lexical changes, but I did feel more aware of the task itself." (T9) and "With the post-editing, I think I edited a bit more freely than before, changing up sentence structure and adding linking words. I was more aware of what I was allowed to do as a translator with decisions." (T7). Four students (T3, T4, T6, T11) remarked that they felt more confident after the training because they understood better how MT works and what types of errors an

MT engine normally makes: "This time around, I felt like I was a lot more consciously aware of the flaws that MT tools can have, and I think that I did a better job fixing them because I knew what to expect going into the text." (T3). One student explicitly mentioned that the training had taught them that terms are not translated consistently in MT output (T6), and one mentioned an increased awareness that the tone and style of the MT may need to be adjusted (T8). Overall, four students explicitly mentioned being more confident after training (T1, T3, T4, T11), and seven explicitly mentioned an increased awareness (T1, T3, T4, T6, T7, T8, T11). While three students had expressed doubts about the correctness of the output and their corrections before training, none of the students explicitly mentioned such doubts after training.

Two students (T3, T7) argued that the training helped them edit more freely. Their explanations show that they especially felt more freedom to make syntactic changes, changing the word order and combining or splitting sentences: "I made a lot of changes to sentence length and structure in particular. In order to avoid repetitive sections, I often merged sentences that would fit together. Additionally, I also split up a lot of sentences that would be simply too long in Dutch." (T3). However, for the current literary excerpts, sentence length and sentence structure played an important role in Hemingway's typical style, and as such, many of these edits were considered errors (overediting) by the reviewer.

Interestingly, the post-training student reflections do not indicate any changes in their opinions about the quality of the raw output or of the final product. Neither does there appear to be any change in their like/dislike for doing MTPE. Both before and after training students were generally satisfied with the quality of the final product and happy to be working with MT as a way to increase their speed and get useful suggestions. As has been explored in the literature before, translators and translation students are not necessarily adversed to technology as long as this technology is useful in their work (Guerberof-Arenas 2013; Koskinen and Ruokonen 2017).

However, four students (T5, T6, T9, T11) still express a clear preference for HT over PE after training because they felt more constrained during the latter ("I took more liberties translating", T11) and it was sometimes hard not to simply accept the output ("I feel like I tend to keep certain phrasings from the MT translation that is [sic] not incorrect but which I would perhaps not have chosen myself.", T6). This confirms findings from earlier studies that PE constrains translators' creativity and takes away from their job satisfaction, but it also suggests that PE training can boost awareness and confidence, and leads some students to edit more freely and creatively.

5. Conclusion

Our goal was to explore the effect MTPE training has on students' creativity. Even though the students reported an increased awareness of how MT works and an increased confidence in doing PE, we found no quantitative evidence

to conclude that training on MTPE significantly affects students' creativity. However, we do see a change both in the quantitative data and in the reflective essays, the students are more willing to try creative shifts after the training, they feel more confident to tackle MT issues, and they feel they have developed their own translation criteria when it comes to translating. This leads them to having more errors in PE, but fewer (albeit not significantly fewer) in HT after the training.

Further, we observe that students show a higher degree of creativity in HT, but significantly fewer errors in PE overall, especially at the start of the training, than when translating on their own. This differs from previous research on creativity with professional translators (Guerberof-Arenas and Toral 2022) where PE caused a higher number of error points. This means that MT might help students work with literary texts when they are still learning or they are insecure about their translations until they acquire enough proficiency through practice in the same way that someone learning a language can benefit from MT at the beginning, but it might cause issues or simply get in the way for a more advanced learner. It also speaks of the high quality of public NMT in this language combination. Finally, HT seems to give students more freedom to create than PE, as we have seen in previous research.

As innumerable studies in Translation Studies show, students do not behave as professionals do and thus it is important to consider this factor when looking at creativity and experimental research with a view to also teaching students how to be creative based on information from professionals.

Perhaps, our goal was too ambitious in the sense that the time elapsed between the beginning of the training and the end was too short for the students to interiorize concepts learnt during the master, but judging from their comments, and although some still prefer to translate on their own (in this sense they share the same opinion as professionals) learning about PE and translation strategies was a positive experience for them. Perhaps a wider time gap between the first and second test could benefit this type of research, but also a more controlled task where students carry out this exercise as part of their course assessment.

For us the most valuable part of this exercise was for students to reflect on how they approach translation and post-editing, and to look at their own changes before and after the training. Because of this, we have now included this as part of the module where students are graded on their reflective essays. Based on the study's findings, we see that training students on the use of MT and PE in literary texts is necessary and appreciated by students, but perhaps more focus on linguistics and error analysis, on self-revision, and on the specific balance between being creative and being accurate is needed. It is true that some of these are acquired through experience (time and practice), but more exercises where students can correct their own texts or the texts of their colleagues using existing error classification might be required to raise this awareness. We suggest, in line with Mellinger (2017, 284), cross-curricular integration of MTPE in practice-oriented and domain-

specific translation modules or courses, instead of only in the translation technology modules. Moreover, we believe that specific courses on creative writing for translators are needed, not only to develop "good" writing, but also to understand how language works from a technical point of view that allows them to identify errors.

Despite its innovative nature and being included within the framework of the CREAMT project and, therefore, using tested instruments, this study is not without limitations. The current study is limited by the number of participating students, the fact that students follow different programmes in two different universities, the limited number of weeks to train students in MT and PE and the fact that the experiment was not graded. Still, the study offers insights into students' creativity when post-editing literary texts and proposes a methodology for students and trainers to reflect on, and re-evaluate, the way in which we teach translation technologies and associated skills. Finally, the open data available can serve to replicate this exercise in other universities and language combinations and, therefore, this first experiment is intended as a starting point and not a destination.

Funding information

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 890697.

Acknowledgements

We would like to thank all the students who provided the data for this experiment as well as the expertise of Leen Van der Broucke who reviewed all the assignments.

References

- **Bayer-Hohenwarter, Gerrit** (2011). "Creative Shifts as a Means of Measuring and Promoting Translational Creativity." *Meta* 56(3), 663–692.

- **Bowker, Lynne** (2002). *Computer-aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press.

- **Bowker, Lynne** (2020). "Machine translation literacy instruction for international business students and business English instructors." *Journal of Business & Finance Librarianship* 25(1-2), 25–43.

- **Bowker, Lynne and Jairo Buitrago Ciro** (2019). *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Leeds: Emerald Group Publishing.

- **CREAMT** (2022) *Creativity and narrative engagement of literary texts translated by translators and neural machine translation* https://cordis.europa.eu/project/id/890697 (last accessed 31.08.2022)

- **Delorme Benites, Alice, Sara Cotelli Kureth, Caroline Lehr, and Elizabeth Steele** (2021). "Machine translation literacy: a panorama of practices at Swiss universities and implications for language teaching". Naouel Zoghlami, Cédric Brudermann, Cédric Sarré, Muriel Grosbois, Linda Bradley, and Sylvie Thouësny (eds) (2021) *CALL and professionalisation: short papers from EUROCALL 2021*.80–87.

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

- **DigiLit** (2023) *Digital literacy in university contexts* https://www.zhaw.ch/en/linguistics/digital-literacy-in-university-contexts-diglit/ (consulted 14.09.2023).

- **Doherty, Stephen and Joss Moorkens** (2013). "Investigating the experience of translation technology labs: pedagogical implications." *Journal of Specialised Translation*, 19, 122–136.

- **Dorst, Aletta G., Susana Valdez, and Heather Bouman** (2022). "Machine translation in the multilingual classroom." *Translation and Translanguaging in Multilingual Contexts* 8(1), 49–66.

- **Dunne, Keiran J.** (2012). "The industrialization of translation: Causes, consequences and challenges." *Translation Spaces* 1, 143–168.

- **Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock** (2023). "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models." *arXiv*. https://arxiv.org/abs/2303.10130 (consulted 08.01.2024).

- **EMT Board and Competence Task-Force** (2022). *European Master's in Translation—EMT Competence Framework 2022*.

- **EMT Expert Group** (2009). *Competences for professional translators, experts in multilingual and multimedia communication*.

- **EMT Expert Group** (2017). *European Master's in Translation—EMT Competence Framework 2017*.

- **Fillmore, Charles J.** (1977). "Scenes-and-frames Semantics." Antonio Zampolli (ed) *Linguistic Structure Processing*. Amsterdam: North Holland Publishing Company, 55-82.

- **Fontanet, Mathilde** (2005). "Temps de créativité en traduction." *Meta* 50(2), 432–447.

- **Guerberof Arenas, Ana, and Joss Moorkens** (2019). "Machine translation and post-editing training as part of a master's programme. *Journal of Specialized Translation*" 31, 217–238.

- **Guerberof-Arenas, Ana** (2013). "What do professional translators think about post-editing?" *The Journal of Specialised Translation* 19, 75–95.

- **Guerberof-Arenas, Ana and Antonio Toral** (2022). "Creativity in translation: Machine translation as a constraint for literary texts." *Translation Spaces* 11(2), 184–212.

- **Heaven, Will Douglas** (2023). "ChatGPT is everywhere. Here's where it came from." *MIT Technology Review*. https://www.technologyreview.com/2023/02/08/1068068/chatgpt-is-everywhere-heres-where-it-came-from/ (consulted 08.01.2024).

- **Kenny, Dorothy (ed)** (2022). *Machine translation for everyone: Empowering users in the age of artificial intelligence*. Berlin: Language Science Press.

- **Kenny, Dorothy and Stephen Doherty** (2014). "Statistical machine translation in the translation curriculum: overcoming obstacles and empowering translators." *The Interpreter and Translator Trainer* 8(2), 276–294.

- **King, Katherine M.** (2019). "Can Google Translate be taught to translate literature? A case for humanists to collaborate in the future of machine translation." *Translation Review* 105: 76–92.

- **Kiraly, Don, Gary Massey, and Sascha Hofmann** (2018). "Beyond teaching : towards co-emergent praxis in translator education." Ahrens, Barbara, Silvia Hansen-Schirra, Monika Krein-Kühle, Michael Schreiber, and Ursula Wienen (eds) (2018). *Translation – Didaktik – Kompetenz*. Berlin: Frank & Timme, 11–64.

- **Koskinen, Kaisa and Minna Ruokonen** (2017). "Love letters or hate mail? Translators' technology acceptance in the light of their emotional narratives." Dorothy Kenny. (ed) (2017) *Human Issues in Translation Technology*. London: Routledge, 8-24.

- **Krüger, Ralph and Janiça Hackenbuchner** (2022). "Outline of a didactic framework for combined data literacy and machine translation teaching." *Current Trends in Translation Teaching and E-Learning*, 375–432.

- **Kussmaul, Paul** (1991). "Creativity in the translation process: Empirical approaches." Kitty M. van Leuven-Zwart and Ton Naaijkens (eds) (1991) *Translation Studies: The State of the Art. Proceedings from the First James S. Holmes Symposium on Translation Studies*Amsterdam: Rodopi, 91-101.

- **Kussmaul, Paul** (1995). *Creativity in Translation. In Training the translator*. Amsterdam: John Benjamins.

- **Loock, Rudy, Sophie Léchauguette, and Benjamin Holt** (2022). "Dealing with the "elephant in the classroom": developing language students' machine translation literacy." *Australian Journal of Applied Linguistics* 5(3), 118–134.

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

- **Mellinger, Christopher D.** (2019). "Metacognition and self-assessment in specialized translation education: task awareness and metacognitive bundling." *Perspectives* 27(4), 604–621.

- **Nitzke, Jean, Anke Tardel, and Silvia Hansen-Schirra** (2019). "Training the modern translator – the acquisition of digital competencies through blended learning." *The Interpreter and Translator Trainer* 13(3), 292–306.

- **Nord, Christiane** (2023). Christiane Nord: "Siempre van a hacer falta traductores humanos." *DUPO - Diario de la Universidad Pablo de Olavide*. https://www.upo.es/diario/entrevista/2023/05/christiane-nord-siempre-van-a-hacer-falta-traductores-humanos/ (consulted 08.01.2024).

- **O'Brien, Sharon** (2002). "Teaching post-editing: A proposal for course content." *Proceedings of the 6th EAMT Workshop "Teaching machine translation" 14-15 November 2002*. 99–106.

- **Panić, Milica** (2019) "DQF-MQM: Beyond Automatic MT Quality Metrics". https://www.taus.net/resources/blog/dqf-mqm-beyond-automatic-mt-quality-metrics (last accessed 14.09.2023).

- **Pym, Anthony** (2014). "Translation Skill-Sets in a Machine-Translation Age." *Meta* 58(3), 487–503.

- **Rothwell, Andrew** (2019). "Tracking Translator Training in Tools and Technologies: Findings of the EMT Survey 2017." *Journal of Specialised Translation* 32, 26-60.

- **Way, Andy** (2019). "Should human translators fear the rise of machine translation?" https://www.adaptcentre.ie/news/should-human-translators-fear-the-riseof-machine-translation (consulted 01.03.2022).

Biographies

Ana Guerberof-Arenas is an Associate Professor at University of Groningen. She was a Marie Skłodowska Curie Research Fellow at the Computational Linguistics group with her CREAMT project that looked at the impact of MT on translation creativity and the reader's experience in the context of literary texts. More recently she has been awarded an ERC Consolidator grant to work on the five-year project INCREC that explores the translation creative process in its intersection with technology in literary and audiovisual translations.

a.guerberof.arenas@rug.nl

https://orcid.org/0000-0001-9820-7074

Susana Valdez is an Assistant Professor in Translation Studies at Leiden University. Before taking up her current position, she had spent 15 years working in the translation industry, and she was an invited lecturer at NOVA School of Social Sciences and Humanities, as well as Lisbon University School of Arts and Humanities (Lisbon, Portugal). Her doctoral thesis (Summa Cum Laude, 2019), conducted in co-tutelle between Lisbon and Ghent universities, dealt with biomedical translation aimed at Portuguese health professionals using keylogging and survey methods. Her research interests include medical translation, the translation process, and reception.

s.valdez@hum.leidenuniv.nl

https://orcid.org/0000-0001-5461-2078

Aletta G. Dorst is an Associate Professor in Translation Studies and English Linguistics at Leiden University. Her research focuses on metaphor variation, metaphor translation, style in translation, literary machine translation, and machine translation literacy. She recently led an NRO Comenius Senior Fellow project on "The value of machine translation in the multilingual academic community" and was the lead researcher for the work package on metaphor identification and translation on the ZonMW Memorabel project "Dementia in metaphors".

a.g.dorst@hum.leidenuniv.nl

https://orcid.org/0000-0003-2520-3157

Notes

[1] The first week of tuition fell in September 2021 and the last in November 2022.
[2] The text is available free of copyright at Project Gutenberg https://www.gutenberg.org/ebooks/61085 (last accessed 14.09.2023).
[3] https://www.deepl.com/nl/translator (consulted 03.09.2021).
[4] The annotation was carried out by the three authors.
[5] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions# (last accessed 14.09.2023).
[6] The detailed list is available here https://github.com/AnaGuerberof/CREAMTTRAINING.
[7] The databases presented here and its analysis are located at https://github.com/AnaGuerberof/CREAMTTRAINING.
[8] These models are used when looking at binary outcomes (there is a CS or not) for repeated or clustered measures (phase, modality) and a random effect (translator).
[9] The databases presented here and their analyses are located at https://github.com/AnaGuerberof/CREAMTTRAINING.
[10] Overdispersion means that the variance is not equal to the mean.
[11] Available at https://www.qualtrics.com/uk/?rid=ip&prevsite=en&newsite=uk&geo=ES&geomatch=uk (last accessed 14.09.2023).
[12] Available at https://github.com/AnaGuerberof/CREAMTTRAINING.