Pre- and Post-Editing for Machine Translation:

Analysing and Explaining the Effects of Data Manipulation in the Context of MT

Table of Contents

1 Introduction

The DataLit^{MT} Framework is a machine translation-specific data literacy framework developed as part of the DataLit^{MT} project. The central dimension of this framework is concerned with *data collection and production* in a machine translation (MT) context. Data collection/ production includes a subdimension called *Data preparation (conversion/manipulation)*, which can be linked to the Professional MT Literacy Framework and here in particular to the two subdimensions *pre-editing for MT* and *MT post-editing* of linguistic MT literacy. In this paper, we examine how textual data can be manipulated in the context of machine translation. This data manipulation can occur both before and after feeding the data into an MT system. In the *pre-editing* stage, a source text is modified in such a way that the MT engine used can produce an output of higher translation quality (Miyata/Fujita 2021:1). In the *post-editing* stage, the MT output is modified in order to correct any errors, and to make the text more readable and comprehensible (ISO 18587 2017:5). Ideally, adequately pre-editing a text for MT reduces the subsequent post-editing effort.

In this paper, we first discuss the theoretical underpinnings of pre-and post-editing and discuss some general rules for pre- and post-editing texts for MT. After laying the theoretical groundwork, we include a practical section in which you can apply the pre- and post-editing rules discussed previously in order to see the effects of pre- and post-editing in practice. For this practical session, we chose a text from the Apple support website explaining the Workout app on an Apple watch. The English source text is already of high quality and reads very well. However, when you translate this source text using the free version of DeepL (or any other MT system of your choice) without applying any pre-editing, you will observe several defects in the output (we use German as the target language in this paper). This shows that a well-written text can still prove difficult to translate for an MT system and that post-editing may still be necessary to optimise the machine-translated target text. In this paper, we show which post-editing steps may be performed in order to optimise the MT output produced from the non-pre-edited source text, and we calculate the translation edit rate (TER) in order to quantify our post-editing effort. In a second step, we pre-edit the source text and feed it again into DeepL. Again, we show how this MT output may be post-edited and measure our post-editing effort by calculating TER. Then, we can compare the effort involved in post-editing the two versions of the source text and see whether pre-editing had any measurable impact on post-editing effort.

2 Pre-editing of source texts

The modern world strives to automate as many manual and intellectual human processes as possible. Of course, this trend doesn't stop when it comes to the translation industry. Constantly improving MT systems that can deliver high-quality output is just one example of how the industry attempts to automate translation processes. But despite all the progress in this field, the output of MT systems is often far from perfect and can be deficient in numerous ways. One way to deal with the problem of inadequate MT output quality is machine translation post-editing or MTPE: automatically translating a text with an MT engine and then modifying, or post-editing, the output according to specific quality criteria (more on this in chapter 3). MTPE is a well-established practice in the translation Industry (Hiraoka/Yamada 2019:64); however, this is not the only way to manipulate textual data in the MT process. Another approach focuses on the source texts that are to be machine-translated: Why not modify, or pre-edit, a source text before feeding it into an MT system?

2.1 What is pre-editing?

Pre-editing follows the idea of modifying the source text "to what the intended [MT] system can properly translate" (Miyata/Fujita 2021:2). For example, an MT system might struggle to translate ambiguous words or a sentence that contains many subordinate clauses but would be able to deliver a good translation if an unambiguous vocabulary were used or if a long and complex sentence were split into at least two shorter sentences. This is not to be confused with plain language writing. Plain language is meant for human readers and uses simple sentences to explain the content in a comprehensible way. In contrast, pre-editing is meant for MT systems – we re-write the text for optimal processing by the MT system. The goal is to obtain higher-quality translations that require a reduced amount of post-editing (Miyata/Fujita 2021:1). Here is what the post-editing norm ISO 18587 (2017) has to say about pre-editing:

> Pre-editing refers to modifying the source language content before machine translation (MT) to facilitate the process, improve raw translation output quality, and therefore reduce the post-editing workload, especially if one document is to be machine-translated into several languages. (ISO 18587 2017:11)

Regarding the powerful but functionally opaque state-of-the-art architecture of *neural machine translation* (NMT), Miyata/Fujita explain that "a deep understanding of what pre-editing is and how it works for black-box NMT is lacking", and that "the impact of pre-editing on black-box MT is unpredictable in nature" (Miyata/Fujita 2021:1f.). However, they also point out that "many studies have demonstrated the effectiveness of pre-editing methods" (Miyata/Fujita 2021:1). Even if we don't know what an MT system actually does with our changes to the source text, we can still see whether these changes result in an improved MT output.

Of course, it is important to consider the type of text we're working with before we decide to pre-edit and machine-translate it. Creative texts that, for instance, use metaphors or play with words that thrive on ambiguity will lose their edge if we edit away everything that makes them special in order to facilitate the text's translation by an MT system. In contrast, technical texts that are written in standardised language and don't rely on linguistic nuances to be understood correctly, e.g., instruction manuals, will probably be less impacted by pre-editing. It is primarily these informative text types that we refer to in the following chapters.

2.2 Pre-editing principles: Negative Translatability Indicators (NTIs) and Controlled Language (CL) rules

If we want to improve the "machine translatability" (O'Brien 2004:1) of a source text, we must be able to recognise the potential problems an MT engine could encounter while translating this text and we need to know how to minimise those problems. In this context, the literature speaks of *Negative Translatability Indicators* (NTI), which can be defined as "linguistic feature[s], either stylistic or grammatical, that [are] known to be problematic for MT" (O'Brien 2005:38). Underwood/Jongejan (2001:363) list five categories of NTIs:

a. structural ambiguity,
b. compounds comprising 3 or more nouns,
c. "sentences" without (finite) verbs,
d. lexical ambiguity, and
e. sentence length (both very long and very short sentences).

O'Brien (2005:39) points out that "the impact of different problems may vary from language pair to language pair". This means that some NTIs might have a stronger impact for certain

language combinations than for others, but they may affect MT in general and should therefore be avoided, if possible.[1]

One way to eliminate NTIs from a source text is to control the language in which the source text is written (O'Brien 2005:38) by following the rules of a so-called controlled language, a "subset of natural languages […] whose grammars and dictionaries have been restricted in order to reduce or eliminate both ambiguity and complexity" (ISO 18587 2017:3). Controlled languages aim "to produce coherent and comprehensible documentation that is easy for [an MT] engine to read" – that means the content is clear and concise, and the wording is simple (Kantan AI 2013).[2] As a result, the style of documents is consistent, texts can be reused, and authoring and translation costs decrease. It should be no surprise that controlled languages are mainly applied in technical documentation. Examples of CLs are the Boeing Technical English (BTE) or the Controlled Automotive Service Language (CASL Project) (Torrejón/Rico 2002:108). Controlled language rules can be adapted to serve as guidelines for pre-editing. But before we explain these, let's have a look at some risks that pre-editing might entail.

2.3 Risks and pitfalls of pre-editing

When pre-editing a source text, we need to make sure that pre-editing does not distort the genre conventions or register requirements of the text type/genre the text belongs to (see also the brief discussion in section 2.1). Of course, this only applies if the source text is actually intended to be used for a particular source-cultural purpose. If the text is to be used merely as a template from which to create multiple target texts, then this text may be pre-edited without such concerns in mind in order to reduce the post-editing effort required for the different target versions. Also, we need to be careful not to introduce any factual or other errors into the source text during pre-editing. This applies in particular to high-risk texts, which may contain security-relevant information, meaning that translation mistakes can "cause danger to life and limb" (Nitzke/Hansen-Schirra 2021:52). If errors are introduced in the pre-editing stage and carried over to the MT output, it falls on the post-editors to identify and correct these mistakes. This may prove difficult because the output of NMT systems is often "deceptively fluent" (Way 2018:164) – meaning that NMT output often reads well and may thus make it difficult to spot any factual errors.

2.4 How to pre-edit

There are different ways to approach pre-editing: bilingual or monolingual pre-editing as well as manual or automated pre-editing. Bilingual pre-editing takes place when the editor modifies "the source text while looking at the translated MT output", requiring the editor to be trained in both the source and the target language. Monolingual pre-editing, on the other hand, only requires the editor to have source language skills (Hiraoka/Yamada 2019:66).

Pre-editing relies heavily on the use of controlled languages when producing the source text and has been done manually for quite some time (Miyata/Fujita 2021:2). In the meantime, tools have been developed that facilitate the pre-editing process (ISO 18578 2017:11), e.g., controlled language checkers such as Acrolinx.[3] For instance, Controlled Language Checkers

[1] Also, note that this list of NTIs was compiled in 2001 and thus long before the introduction of NMT (around 2015). In this chapter, we discuss a rather extensive set of potential NTIs and associated pre-editing rules. Keep in mind, however, that NMT often has less problems than previous MT architectures when translating these NTIs.

[2] Note, however, that Marzouk/Hansen-Schirra (2019) did not find any positive impact of using controlled language (which can be considered "a special case of pre-editing", ISO 18587 2017:11) on the output of the NMT system investigated by the authors (which was Google Translate).

[3] Also, large language models such as ChatGPT may also be able to pre-edit text for MT, although their ability (and reliability) to do so has not yet been tested in large-scale studies.

can "automatically [break] each sentence into brief natural phrases, which can then be manually translated where necessary" (Miyata/Fujita 2021:2). These tools allow users to filter the rules they want to apply. Pre-editing in a wider sense, or linguistic pre-editing, means that the source-language content is checked according to controlled language rules, which improves the machine translatability of the source text. In a narrower sense, pre-editing merely involves checking the source text for correct spelling and formatting, and ensuring that text strings that should not be translated are tagged accordingly (cf. ISO 18587 2017:11).

Miyata and Fujita (2021:7) propose a number of strategies for pre-editing a source text:

a. information addition: make implicit information explicit to clarify the meaning of the text,

b. use of clear relations: change the sentence structure and use explicit connective markers to clarify relations between words, phrases, and clauses,

c. use of narrower sense: replace general words with more specific ones that convey a certain meaning, and

d. normalisation: use standardised expressions, style, and notation.

These strategies are also a common denominator of many controlled languages and, as such, they also influence technical writing and pre-editing rules. When it comes to technical writing, many sets of rules and extensive guidelines have been established by various companies and organisations, for example the *Gesellschaft für Technische Kommunikation (tekom)* in Germany. For our purposes, we don't need to have hundreds of pages filled with rules and how-tos. Rather, we focus on a range of pre-editing rules that have shown to be effective for improving the quality of MT output across different languages. For specific language combinations, text types, and genres, only a subset of these rules may be applicable. The rules are compiled from various sources that propose best-practices in the field of controlled language and pre-editing rules, namely Drewer/Ziegler (2014), Kantan AI (2013), and Muegge (2008). We have also included the findings from recent pre-editing studies carried out by Miyata/Fujita (2021) and Hiraoka/Yamada (2019). In summary, the rules are as follows:

Rule 1: Use correct and consistent spelling and punctuation

Spelling and punctuation should always be consistent. For English source texts, e.g., a pre-editor might need to check for the correct and consistent use of American or British English and make sure that punctuation (such as commas) is used correctly and consistently.

Example: "optimize" (AE) vs. "optimise" (BE)

Example: not: "Alternatively to the search column you can use the form."

but: "Alternatively to the search column, you can use the form."

Rule 2: Keep it short and concise

Split up very long sentences (with more than 60 words). Avoid conjunctions and more than one clause when possible. Convey only one idea per sentence.

Example: not: "The execution of the saving process is triggered by pressing the Save button."

but: "To save the file, select Save."

Rule 3: Be explicit

Re-write sentences if they are not explicit enough. It is okay for the revised sentence to be longer than the original.

Example: not: “The stock market was closed on the first.”

but: “The stock market was closed on the first of May due to a holiday.”

Rule 4: Use complete sentences and simple grammar

Make sure that sentences are grammatically complete: beginning with a capital letter, having at least one main clause, and ending punctuation. Identify complex grammatical structures and break them up into separate sentences if possible.

Example: not: “Continue installing software?”

but: “Do you wish to continue installing the software?”

Example: not: “You, in your texts, to show that you can organise your thoughts, should use a simple sentence structure.”

but: “Show that you can organise your thoughts by using a simple sentence structure in your texts.”

Rule 5: Use the active voice

If possible, re-write sentences in the active voice to avoid ambiguity.

Example: not: “The image is shown in the preview window in a reduced quality.”

but: “The preview window shows the image in a reduced quality.”

Rule 6: Be consistent in your use of syntax and terminology

If several sentences convey the same idea, re-write them as the same sentence and use that sentence consistently in the document. Make sure the terminology is consistent as well.

Example: not: “Lift the scanner from the shipping box. Afterwards, you can discard all packaging material.”

but: “Remove the scanner from the carton. Remove the plastic wrapping.”

Example: not: “wind turbine, wind-powered generator”

but: “wind turbine”

Rule 7: Use words from a general dictionary

Replace unusual words with more usual synonyms if they convey the same meaning.

Example: not: “Turn on your data processor.”

but: “Turn on your computer.”

Rule 8: Repeat nouns instead of using pronouns

Be explicit when referring to a noun. Repeat the noun instead of using a replacing pronoun.

Example: not: “The car drove past the house. It was green. It was blue.”

but: “The car drove past the house. The house was green. The car was blue.”

Rule 9: Specify nouns using "the"

Avoid ambiguity (especially with gerunds) and ellipses.

Example: not: "Train KantanMT engine."

but: "Train the KantanMT engine."

Rule 10: Remove needless words that don't contribute to the meaning of a sentence.

Example: not: "Train the KantanMT Machine Translation engine."

but: "Train the KantanMT engine."

Rule 11: Avoid clichés and colloquial phrases

MT systems may not convey the correct meaning of colloquial phrases (this will depend largely on their training data) and the meaning may not make sense to international users.

Example: not: "It is a piece of cake."

but: "It is easy."

Rule 12: Check the content for logical order and completeness

The source text should be structured logically and be complete in terms of content. If information is missing, the pre-editor can add this information to the text (which may require corresponding subject-matter knowledge).

Example: not: "The local office continues the search for the missing hikers."

but: "The local police office continues the search for the missing hikers."

Rule 13: Mask sensitive data

Depending on the MT system, sensitive data should be protected by masking it either manually or with an encryption key. See, for example, the Trados Project Anonymizer.

3 Post-editing of MT output

As discussed briefly in chapter 1, post-editing a machine-translated text can ensure an adequate quality of the MT output. Simply put, post-editing describes all editing steps that are performed after a text has been automatically translated. State-of-the-art MT systems can produce translations of sometimes impressive quality, but they still regularly produce deficient translations or outright translation errors. Depending on the intended use of machine-translated texts, they should always be post-edited by a qualified professional translator (Expert-in-the-Loop production model).

3.1 What is post-editing?

Generally speaking, there are two types of post-editing: *full post-editing* and *light post-editing*. Which of these two types to choose will depend on the purpose of the translated content. Is the translation intended for internal use only and not for publication? In this case, light post-editing is just fine: While the content needs to be comprehensible and accurate, language style and other more granular aspects of translation aren't that important. Yamada (2014:2) describes the goal of light post-editing as "fit for use". TAUS stresses the "good enough" quality a lightly post-edited translation should have (Nitzke/Hansen-Schirra 2021:30). In other words, "use as much of the MT output as possible" (ISO 18587 2017:8).

In contrast, when translating legally binding documents, high-risk texts, highly visible marketing texts or other texts that have to be flawless both in terms of style and content, full post-editing will usually be the right choice. The post-editing norm ISO 18587 states that – while as much MT output as possible should be used – the result of full post-editing should be "a product comparable to a product obtained by human translation" (ISO 18587 2017:2). We will have a look at these guidelines in chapter 3.3.

Modifying a text within the post-editing stage of the translation process involves different types of effort: temporal, technical, and cognitive (Krings 2001). The temporal effort refers to the time spent editing; the technical effort refers to the number of keyboard strokes and/or mouse clicks required to insert, delete, or move around parts of text; and lastly, the cognitive effort refers to the mental effort required to identify deficiencies or errors in the source text and to plan corrections (Lacruz 2017:386). Cognitive effort is certainly the most interesting effort type for gaining insights into the post-editing process; however, in contrast to temporal and technical effort, it is quite hard to measure.

In our practical section 4 below, we will approximate technical post-editing effort by calculating the Translation Edit Rate (TER) in order to compare post-editing effort with and without pre-editing the source text. "TER measures the number of edits that are necessary to go from the raw MT output to a final edited version" (Luongo 2016). It can be calculated based on the number of words or the number of characters in one sentence (in our examples below, we use word-based TER). Essentially, TER reflects the quality of MT output by dividing the sum of all edits, i.e. additions, deletions, substitutions, and changes in position (shifts), by the number of words in the final post-edited version of the text, i.e. the reference translation (see figure 1 below). The higher the TER value, the more edits have been carried out (cf. Snover et al. 2006:3).

$$\text{TER} = \frac{additions + deletions + substitutions + changes\ in\ position}{length\ of\ reference\ translation}$$

Fig. 1: How to calculate the TER

3.2 Risks and pitfalls of post-editing

Translation projects should always be assessed in the planning stage in order to decide whether we want to use MT (including pre- and post-editing) or human translation, and this decision-making process has to consider different aspects (for more information on this, see Nitzke et al 2019:242–246). Since every project brings different kinds of risks and pitfalls, communicating with the client is key to this question. However, if you have already decided to go ahead with MT, what do you need to consider?

Whether or not a source text has been pre-edited, if it is a high-risk text with highly sensitive content, translation mistakes can "cause danger to life and limb" (Nitzke/Hansen-Schirra 2021:52, see also section 2.3 above). Post-editors will have to be extra careful when editing these kinds of texts to make sure the translation conveys the correct meaning.

If your document contains sensitive data that has been masked by a pre-editor before using MT (see rule 13 in section 2.4), post-editors will have to demask the data correctly. If the pre-editor masked the data manually, the post-editor would also have to demask it manually by copying the relevant data from the source text. If the pre-editor used an encryption key, the post-editor would have to use the corresponding decryption key to decrypt the data (cf. Satori Cyber Ltd n. d.).

3.3 How to post-edit

ISO 18587 (2017) includes guidelines on which aspects to pay attention to when post-editing. Since most content in the translation industry is meant to be published and therefore needs to be flawless, we will limit ourselves to the requirements of full post-editing and disregard light post-editing here.[4] The ISO standard explains that the results of full post-editing should be "accurate, comprehensible and stylistically adequate, with correct syntax, grammar and punctuation" (ISO 18587 2017:8).

The requirements below are adapted from ISO 18587 (2017:8) and describe full post-editing as follows:

1. Double-check the source text if you suspect the machine has added or omitted content.
2. Use the appropriate register for the target audience.

   Example: In some languages, the audience can be addressed formally or informally. Choose one form and use it consistently throughout the text.
3. Restructure sentences if the meaning is unclear. A sentence can be correct content-wise but have an unwieldy structure. If this is the case, try to keep it short and concise.

   Example: Split long sentences.
4. Use correct grammar, syntax, and semantics of the target language.

   Example: If a term is ambiguous and could refer to more than one concept, use a synonym that is not ambiguous.
5. Keep your client's requirements in mind, such as consistent terminology or corporate language.

   Example: Are customers addressed formally or informally?
6. Be consistent with spelling, punctuation, and hyphenation, especially regarding language variations.

   Example: "optimize" (AE) vs. "optimise" (BE)
7. Comply with genre conventions/register requirements.

   Example: Don't use literary stylistic elements if the text is concerned with a technical subject matter.
8. Apply appropriate formatting.

   Example: Check if any bold text in the target text was also bold in the source text and adjust if necessary.

4 Practical application

Now it's your turn. To help you practice pre- and post-editing and to observe potential positive effects of pre-editing on post-editing effort, you'll now work on an instruction text from the Apple support website explaining the use of the Workout app on the Apple Watch (13 July 2022). We shortened the text in order to illustrate relevant textual aspects relevant for pre- and

[4] More information on light post-editing can be found in Annex B of ISO 18587 (2017:10).

post-editing. Below, you'll find a description of the individual tasks, the original source text as well as sample solutions for pre-and post-editing this text.

4.1 Tasks

Task 1: Post-editing the machine-translated version of the source text where no pre-editing was applied

a) Translate the source text "Use the Workout app on your Apple Watch" (see section 4.2 below) with an MT system of your choice (in this paper, we used DeepL).

b) Look at the output (our DeepL solution is provided in section 4.4 below) and post-edit the target text until you're happy with the result, keeping in mind the post-editing rules that we explained in chapter 3 of our paper. We provide a sample solution in section 4.6 below.

c) Measure your post-editing effort by calculating the word-based translation edit rate (TER) using this notebook.

Task 2: Pre-edit the source text, post-edit its machine-translated version again and compare the differences in post-editing effort between tasks 1 and 2

a) Take a closer look at the source text from task 1a and modify it according to the pre-editing rules discussed in chapter 2 of our paper (we provide a sample solution in section 4.3 below). Check the output from task 1b to see which parts of the source text may have to be edited in order to obtain a better MT output.

Note: The modified source text is meant to help you obtain a higher-quality MT output. It will not be published in place of the original source text but is only used in MT. So you can focus on pre-editing the source text with a view to improving target text quality without having to worry about deteriorating source text quality during pre-editing.

b) Translate the pre-edited source text with the same MT system you used for task 1 (again, in this paper, we used DeepL; you can find the output in section 4.5 below).

c) Post-edit the text again so that it is identical to your post-edited text from task 1 (see section 4.6 below). Again, measure the post-editing effort by calculating the word-based TER using this notebook.

d) Compare the TER values from tasks 1 and 2. Did pre-editing the source text result in a lower TER value (indicating a lower post-editing effort)?

4.2 Original source text

Use the Workout app on your Apple Watch

Get progress updates while you work out and know when you reach your goal.

Start a workout

1. Open the Workout app.
2. Find the workout that best matches what you're doing. Learn more about each workout type.
3. To set a goal, tap the More button next to the workout that you want to do. To skip setting a goal, tap the workout.
4. Wait for the three-second countdown. To skip the countdown, tap the screen.
5. Work out. To add another type of workout without ending your session, open the Workout app, swipe right, then tap the New button.

4.3 Pre-edited source text

The text below is a sample solution for pre-editing our source text. The pre-edited parts are highlighted in green.

Using the Workout app on your Apple Watch

You get updates on your progress while you work out and get notified when you reach your goal.

Start workout:

1. Open the Workout app.
2. Select the workout that best matches your activity. Learn more about each workout type.
3. To set a goal, tap the More button next to the workout that you want to do. If you don't want to set a goal, tap the workout.
4. Wait for the countdown of three seconds. To skip the countdown, tap the screen.
5. Start the workout. To add another type of workout without ending your active workout, open the Workout app, swipe right, then tap the New button.

4.4 Original DeepL output

Verwenden Sie die Workout-App auf Ihrer Apple Watch

Erhalte Fortschrittsaktualisierungen, während du trainierst, und erfahre, wann du dein Ziel erreicht hast.

Starten Sie ein Workout

1. Öffnen Sie dieWorkout-App.
2. Suchen Sie das Training, das am besten zu Ihrem Vorhaben passt. Erfahren Sie mehr über jeden Trainingstyp.
3. Um ein Ziel festzulegen, tippen Sie auf die Schaltfläche Mehr neben dem Training, das Sie durchführen möchten. Um das Setzen eines Ziels zu überspringen, tippen Sie auf das Training.
4. Warten Sie den dreisekündigen Countdown ab. Um den Countdown zu überspringen, tippen Sie auf den Bildschirm.
5. Trainieren Sie. Um eine andere Art von Training hinzuzufügen, ohne Ihre Sitzung zu beenden, öffnen Sie die Workout-App, wischen Sie nach rechts und tippen Sie dann auf die Schaltfläche Neu.

4.5 DeepL output after pre-editing the source text

Verwenden der Workout-App auf Ihrer Apple Watch

Sie erhalten Updates zu Ihrem Fortschritt, während Sie trainieren, und werden benachrichtigt, wenn Sie Ihr Ziel erreichen.

Workout starten:

1. Öffnen Sie die Workout-App.
2. Wählen Sie das Training, das am besten zu Ihrer Aktivität passt. Erfahren Sie mehr über jeden Trainingstyp.
3. Um ein Ziel festzulegen, tippen Sie auf die Schaltfläche Mehr neben dem Training, das Sie durchführen möchten. Wenn Sie kein Ziel festlegen möchten, tippen Sie auf das Training.
4. Warten Sie den Countdown von drei Sekunden ab. Um den Countdown zu überspringen, tippen Sie auf den Bildschirm.

5. Starten Sie das Training. Um eine andere Art von Training hinzuzufügen, ohne Ihr aktives Training zu beenden, öffnen Sie die Workout-App, wischen Sie nach rechts und tippen Sie dann auf die Schaltfläche Neu.

4.6 Post-edited DeepL output

The text below is a sample solution for post-editing the MT output produced by machine-translating the original and the pre-edited source texts (for reasons of comparability, we use the same post-edited version for both MT outputs). The post-edited parts are highlighted in green.

Trainings-App auf der Apple Watch verwenden

Du erhältst während des Trainings Updates zu deinem Fortschritt und wirst darauf aufmerksam gemacht, wenn du dein Ziel erreicht hast.

Training starten

1. Öffne die Trainings-App.
2. Wähle das Training, das deiner Aktivität am ehesten entspricht. Hier erfährst du mehr über die einzelnen Trainingsarten.
3. Um ein Ziel festzulegen, tippe neben dem Training, das du absolvieren möchtest, auf die Taste für weitere Optionen. Wenn du kein Ziel festlegen möchtest, tippe auf das Training.
4. Warte den Countdown von 3 Sekunden ab. Du kannst den Countdown überspringen, indem du auf den Bildschirm tippst.
5. Beginne mit dem Training. Wenn du eine neue Trainingsart hinzufügen möchtest, ohne das aktuelle Training zu beenden, öffne die Trainings-App, streiche nach rechts, und tippe auf die Neu-Taste.

4.7 Some changes performed during pre- and post-editing

Below, we discuss some pre- and post-editing operations performed on the texts listed in sections 4.3 and 4.6 above.

Form of address

As you can see, the DeepL output is inconsistent in its form of address, mostly addressing the German target text readers in a formal way (*Sie*) but also sometimes opting for the informal *Du*. Addressing readers in a consistent way is very important (see post-editing rule 2 in section 3.3) and, in line with Apple's corporate language, we used the informal *Du* consistently in the post-edited target text. Since the English equivalent of *Du* and *Sie* is *you* in both cases, you can't control the level of politeness in the German MT output by pre-editing the English source text, so this is an aspect that you always have to keep in mind and adjust if necessary when post-editing English-German translations.

Terminological consistency

DeepL switches between *Training* and *Workout* when translating the English term *workout* and between *Taste* and *Schaltfläche* when translating the English term *button*. Terminological inconsistencies in the source text can (and should) be fixed during pre-editing (see pre-editing rule 6 in section 2.4). However, in this case, the source text uses both *workout* and *button* consistently and still there is terminological inconsistency in the MT output. Again, this is an aspect that you should keep in mind and pay particular attention to during post-editing (see post-editing rule 5 in section 3.3).

Vague/implicit vs. precise/explicit language

Sometimes, a source text can express information in a vague or implicit way (relying on the context to provide a precise interpretation), which makes it difficult for an MT system to produce a correct translation. For example, in the source text string *the workout that best matches what you're doing* the part *what you're doing* is semantically vague and prompts DeepL to produce the contextually inadequate translation *Ihrem Vorhaben*. During post-editing, this was then changed to the contextually adequate translation *deiner Aktivität*. Vague/implicit source text passages are well-suited for pre-editing (see pre-editing rule 3 in section 2.4) in order to help the MT system produce a contextually adequate translation.

Imperative vs. indicative mood

English source texts can be written in imperative mood where the indicative mood may be more adequate in German. For example, DeepL translated the English sentence *Get progress updates while you work out and know when you reach your goal* as *Erhalte Fortschrittsaktualisierungen, während du trainierst, und erfahre, wann du dein Ziel erreicht hast*, preserving the imperative mood in the German source text. During post-editing, we changed this to *Du erhältst während des Trainings Fortschrittsberichte und wirst darauf aufmerksam gemacht, wenn du dein Ziel erreicht hast*, thereby rendering the sentence in indicative mood. If the English source text allows both the imperative and the indicative mood in instances where one of the two modes is preferred in German, the mood can be adjusted accordingly during pre-editing.[5]

Compounds vs. prepositional word groups

Both English and German allow for easy compounding of words. However, compounds may sometimes be cumbersome to read, and expanding them into prepositional word groups may improve readability. In our example text, DeepL translated the English compound *progress updates* as *Fortschrittsaktualisierungen*, which we expanded into the word group *Updates zu Ihrem Fortschritt* during post-editing. If the source text allows both compounding and using prepositional word groups for certain multi-element terms, these instances may be pre-edited according to the preferred solution in the German target text (MT systems tend to translate compounds as compounds and word groups as word groups, although this will also depend on a system's training data).

4.8 Comparing post-editing effort

If you calculated TER scores for the MT output produced by machine-translating the *original source text* and the target text post-edited based on this MT output (TER score 1) and for the MT output produced by machine-translating the *pre-edited source text* and the target text post-edited based on this MT output (TER score 2), you should obtain a TER score 1 of 73.81 and a TER score 2 of 67.46.[6] As TER is a distance measure, a higher TER value indicates a higher post-editing effort (and hence a lower MT quality) and a lower TER value indicates a lower post-editing effort (and hence a higher MT quality). With respect to our example, the two TER values indicate that post-editing the MT output based on the original source text entailed a higher post-editing effort (TER score of 73.81) than post-editing the MT output based on the pre-edited source text (TER score of 67.46). In other words, in our example, pre-editing the source text indeed had a measurable positive impact on subsequent post-editing effort. If you consider that, in the professional translation industry, it is often the case that one source text is

[5] Since this is a very language combination-specific issue, we do not list any corresponding pre-and post-editing rules in sections 2.4 and 3.3.

[6] Of course, you'll obtain these exact values only if you adhered exactly to our sample solutions above. If you performed different changes during pre- and post-editing, you'll also obtain different TER values.

translated into multiple (let's say 10) target languages, it may indeed make sense to accept the effort of pre-editing one source text if this may reduce the post-editing effort of ten translators.

5 Conclusion

In this paper, we discussed data manipulation in the context of MT by means of pre- and post-editing. We discussed the theoretical underpinnings of pre- and post-editing as well as some high-level pre- and post-editing rules and provided a practical example, which showed that pre-editing a source text may indeed decrease subsequent post-editing effort. Pre- and post-editing texts are important linguistic tasks in modern MT-assisted translation production workflows (forming part of linguistic MT literacy) and being able to measure the potential benefits of pre-editing for the subsequent post-editing stage is an important process-related aspect of economic MT literacy.

References

Drewer, Petra/Ziegler, Wolfgang (2014[2]): *Technische Dokumentation. Eine Einführung in die übersetzungsgerechte Texterstellung und in das Content-Management.* Würzburg: Vogel.

Hiraoka, Yusuke/Yamada, Masaru (2019): "Pre-editing plus neural machine translation for subtitling: Effective pre-editing rules for subtitling of TED talks." In: Forcada, Mikel/Way, Andy/Tinsley, John/Shterionov, Dimitar/Rico, Celia/Gaspari, Federico (Eds.): *Proceedings of the machine translation summit XVII: Translator, project and user tracks*. Dublin: European Association for Machine Translation, 64–72. https://aclanthology.org/W19-6710 (19 February 2023).

ISO 18587 (2017): *Translation services — Post-editing of machine translation output — Requirements*. Geneva: ISO Copyright Office.

Kantan AI (2013): *How to write for machine translation*. *Kantan AI*. https://kantanmtblog.com/2013/07/03/how-to-write-for-mt (18 February 2023).

Krings, Hans P. (2001): *Repairing texts. Empirical investigations of machine translation post-editing processes*. Kent: University Press.

Krüger, Ralph/Hackenbuchner, Janiça (2022): A didactic framework for combined data literacy and machine translation literacy teaching for translation students. *Current Trends in Translation Teaching and Learning E*, 375–432. http://dx.doi.org/10.51287/cttl202211.

Lacruz, Isabel (2017): Cognitive effort in translation, editing, and post-editing. In: Schwieter, John W./Ferreira, Aline (Eds.): *The handbook of translation and cognition*. New Jersey: John Wiley & Sons, 386–401. http://dx.doi.org/10.1002/9781119241485.ch21.

Luongo, Sharon (2016): Translation edit rate T.E.R. *RWS Community*. https://community.rws.com/product-groups/linguistic-ai/b/weblog/posts/translation-edit-rate-t-e-r (18 February 2023).

Marzouk, Shaimaa/Hansen-Schirra, Silvia (2019): Evaluation of the impact of controlled language on neural machine translation compared to other MT architectures. *Machine Translation* 33(1–2), 179–203. https://doi.org/10.1007/s10590-019-09233-w.

Miyata, Rei/Fujita, Atsushi (2021): Understanding pre-editing for black-box neural machine translation. *arXiv*. https://doi.org/10.48550/arXiv.2102.02955 (17 February 2023).

Muegge, Uwe (2008): *Controlled Language: Rules for Machine Translation.* http://www.muegge.cc/controlled-language.htm (17 June 2022).

Nitzke, Jean/Hansen-Schirra, Silvia (2021): *A short guide to post-editing*. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.5646896.

Nitzke, Jean/Hansen-Schirra, Silvia/Canfora, Carmen (2019): Risk management and post-editing competence, In: *Journal of Specialised Translation* 31, 239–259. https://jostrans.org/issue31/art_nitzke.php (17 February 2023).

O'Brien, Sharon (2004): Machine translatability and post-editing effort. How do they relate. In: *Proceedings of translating and the computer 26*. Aslib, 1–31. https://aclanthology.org/2004.tc-1.3.

O'Brien, Sharon (2005): Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability. In: *Machine Translation* 19(1), 37–58. https://doi.org/10.1007/s10590-005-2467-1.

Satori Cyber Ltd (n. d.): Data Masking: 8 Techniques and How to Implement Them Successfully. *Satori Cyber*. https://satoricyber.com/data-masking/data-masking-8-techniques-and-how-to-implement-them-successfully (18 February 2023).

Snover, Matthew/Dorr, Bonnie/Schwartz, Rich/Micciulla, Linnea/Makhoul, John (2006): A study of translation edit rate with targeted human annotation. In: *Proceedings of the 7th conference of the association for machine translation in the Americas: Technical papers*. Cambridge: Association for Machine Translation in the Americas, 223–231. https://aclanthology.org/2006.amta-papers.25 (18 February 2023).

Torrejón, Enrique/Rico, Celia (2002): Controlled translation: A new teaching scenario tailor-made for the translation industry. In: *Proceedings of the 6th EAMT workshop: Teaching machine translation*. Manchester: European Association for Machine Translation, 107–116. https://aclanthology.org/2002.eamt-1.12 (18 February 2023).

Underwood, Nancy L./Jongejan, Bart (2001): Translatability checker: a tool to help decide whether to use MT. In: Maegaard, Bente (Ed.): *Proceedings of the Machine Translation*

Summit VIII. Santiago de Compostela, 363–368 https://aclanthology.org/2001.mtsummit-papers.65 (18 February 2023).

Way, Andy (2018): Quality expectations of machine translation. In: Moorkens, Joss/Castilho, Sheila/Gaspari, Federico/Doherty, Stephen (Eds.): *Translation quality assessment. From principles to practice*. Cham: Springer, 159–178. https://doi.org/10.1007/978-3-319-91241-7_8.

Yamada, Masaru (2014): Can college students be post-editors? An investigation into employing language learners in machine translation plus post-editing settings. In: *Machine Translation* 29(1), p. 49–67. https://doi.org/10.1007/s10590-014-9167-7.