Chapter 4

Selecting and preparing texts for machine translation: Pre-editing and writing for a global audience

Pilar Sánchez-Gijón
Universitat Autònoma de Barcelona

Dorothy Kenny
Dublin City University

Neural machine translation (NMT) is providing more and more fluent translations with fewer errors than previous technologies. Consequently, NMT is becoming a real tool for speeding up translation in many language pairs. However, obtaining the best raw MT output possible in each of the target languages and making texts suitable for each of the target audiences depends not only on the quality of the MT system but also on the appropriateness of the source text. This chapter deals with the concept of pre-editing, the editing of source texts to make them more suitable for both machine translation and a global target audience.

1 Introduction

Put simply, *pre-editing* involves rewriting parts of source texts in a way that is supposed to ensure better quality outputs when those texts are translated by machine.[1] It may involve applying a formal set of rules, sometimes called *controlled*

[1] As discussed in Rossi & Carré (2022 [this volume]), quality is not a fixed concept; rather, judgments about quality depend on a whole host of factors, including the intended purpose of a translation. For a detailed discussion of this highly mutable concept, see Drugan (2013) and Castilho et al. (2018).

Pilar Sánchez-Gijón & Dorothy Kenny. 2022. Selecting and preparing texts for machine translation: Pre-editing and writing for a global audience. In Dorothy Kenny (ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence*, 81–103. Berlin: Language Science Press. DOI: 10.5281/zenodo.6759980

language rules, which stipulate the specific words or structures that are allowed in a text, and prohibit others (see, for example, O'Brien 2003). Alternatively, it can involve applying a short list of simple "fixes" to a text, to correct wrong spellings, or impose standard punctuation, for example. Depending on the context, it might involve both of the above. Whatever the case, its main purpose, as understood here, is to improve the chances of getting a better quality target text once the source text has been machine translated. In cases where a source text is to be translated into multiple target languages, the benefits of pre-editing should, in theory, be observed over and over again in each of the target language texts. It is thus traditionally recommended in multilingual translation workflows.

Another way to ensure that a text is translatable is to write it that way in the first place. Writers whose work will ultimately be translated into multiple languages are thus often asked to write with a global audience in mind. As well as applying principles of "clear writing", they are asked, for example, to avoid references that may not be easily understood in cultures other than their own. This applies also to writers whose work will be read in the original language by international readers who are not native speakers of that language.

Given their similar aims, it is not surprising that there is often overlap between pre-editing rules, controlled languages and guidelines for clear writing or writing for global audiences. In this chapter, we give an overview of the kind of guidance commonly encountered in such sources, without attempting to be exhaustive. The reader must also remember that such guidance is always language specific: advice about the use of tense forms, for example, applies only to languages that have grammatical tense. Many do not. Guidance can also be language-*pair* specific or specific to a particular machine translation (MT) type or engine. A construction that caused problems in rule-based MT (RBMT) may no longer be an issue in neural MT (NMT), or it might be associated with errors in a neural engine trained on legal texts but not one trained on medical texts. In the case of writing with MT in mind, what turns out to be useful advice thus depends heavily on the context.

The advent of NMT, in particular, has made us rethink the usefulness of advice on pre-editing and controlled writing (see Marzouk & Hansen-Schirra 2019 and §2 below), but for much of the history of MT, pre-editing helped ensure the success of the technology. A good knowledge of MT made it possible to predict those aspects of the source language or the source text that would likely generate errors in translations produced by a given MT system, whether rule-based or statistical. However, one of the aspects that characterize NMT is precisely its lack of systematic error: it can be difficult to predict with any certainty what type of error will occur, and so attempting to pre-empt particular errors may seem ill-advised. The substantial improvement in the fluency and adequacy of translations obtained through NMT might also suggest that steps taken to improve output are an unnecessary luxury. In the context of NMT, it might seem, in other words, that pre-editing is, a priori, redundant. Improvements in MT do not diminish the benefits of all types of pre-editing, however. While some traditional pre-editing approaches may no longer be relevant, as will be discussed below, others become essential, especially if pre-editing is included in a translation pipeline in which there is no post-editing or in which "good enough" post-editing (see O'Brien 2022: §2 [this volume]) is performed. What is more, while improvements in quality certainly mean that the translation problems which used to characterize MT have been reduced to a great extent, errors have not been completely eliminated, as we will see below, and new errors are emerging for which MT has not been evaluated so far. These are linked to the nature of the translation commission (see below), the function of the source text and the assumed intention of its author, and it is in these cases that pre-editing continues to play a role in optimizing the use and increasing the effectiveness of MT.

The rest of this chapter starts out by discussing the background and uses of pre-editing in the more recent past and in current uses of NMT. The chapter goes on to describe the strategies involved in selecting texts for use with MT, and with the influence of English as a source language on machine translated texts. It then presents the case for writing for a global audience to start with. The chapter concludes by presenting common pre-editing guidelines, as well as the resources and tools used in this task.

2 Pre-editing and NMT

In the past, when rule-based systems produced obvious, and often systematic, errors in adequacy and fluency (see Rossi & Carré 2022 [this volume]), pre-editing was often necessary to get the best out of MT. Even after the transition to statistical MT (SMT), researchers still found pre-editing to be useful. Seretan et al. (2014), for example, working with the language pairs English-French, English-German and French-English, found that appropriate pre-editing led to quality improvements in the MT of user-generated content in both the technical and health domains. In a related study, also using SMT, Gerlach (2015) found that pre-editing English source texts resulted in faster post-editing of machine translations into French, although the overall impact on productivity in the extended workflow was less clear. Likewise, Miyata & Fujita (2017) found that pre-editing Japanese texts resulted in better translations into English, Chinese and Korean, confirming, in their view, the particular usefulness of pre-editing in multilingual SMT settings (ibid.: 54). The same researchers subsequently studied the influence of pre-editing on the output of two NMT systems, but this time found that there was very little correlation between the amount of pre-editing done and the amount of post-editing that was needed after pre-edited texts were machine translated from Japanese into English, Chinese or Korean (Miyata & Fujita 2021). Miyata and Fujita (ibid.) also looked at the effect of different types of pre-edits, and found that the edits that had been traditionally recommended in the context of MT were less frequently encountered in NMT workflows:

> Contrary to the acknowledged practices of pre-editing, the operation of making source sentences shorter and simpler was not frequently observed. Rather, it is more important to make the content, syntactic relations, and word senses clearer and more explicit, even if the ST becomes longer. (Miyata & Fujita 2021: 1547).

Other studies suggest that pre-editing is simply not an effective strategy with NMT systems. Marzouk & Hansen-Schirra (2019), for example, found that pre-edits improved the performance of an RBMT, an SMT, and a hybrid MT system, in the context of German-to-English technical translation, but they did not improve the performance of the NMT system they tested.[2] Among the few studies that are enthusiastic about pre-editing in the context of NMT is that by Hiraoka & Yamada (2019). They applied just three pre-editing rules to Japanese TED Talk subtitles, namely:

- fill in missing punctuation,
- fill in missing grammatical subjects and/or objects,[3] and
- write out proper nouns in the target language.

According to Hiraoka and Yamada (ibid.), the implementation of these three edits improved the translation into English of the subtitles using an off-the-shelf NMT system. In some rare cases, however, it resulted in dis-improvements in the MT output.

[2] Like other authors, Marzouk & Hansen-Schirra (2019) are careful to point out that their research is based on so-called *black-box* systems, that is, off-the-shelf systems whose internal workings cannot be scrutinized by the analyst.

[3] Japanese, like Spanish, is a "pro-drop" language (see Kenny 2022: §1 [this volume]), meaning that certain pronouns can be omitted without impeding comprehension. In Japanese, these can be either subject or object pronouns.

Given the lack of clear research evidence to support the use of pre-editing in NMT workflows, industrial users of NMT are best advised to test the effects of pre-edits carefully before promoting their use in production environments. As indicated in the Introduction to this chapter, they may find that certain edits are useful only for particular language pairs, given particular genres and particular NMT engines and the training data they are based on.

3 Genre and domain-based advice

From the point of view of professional translators (and the translation industry), the use of MT on a regular and integrated basis within translation projects is associated with specific text *genres* and *domains* (see Kenny 2022: §1 [this volume]). In the past, the focus on genres and domains that used predictable, and sometimes repetitive or restricted words and structures, meant that controlled language approaches made sense: in contexts where a quick and cost-effective translation was needed (for example, in the case of in-house technical documentation that was not available to the general public), the words and structures used in source texts were controlled to ensure the success of the MT. In the era of data-driven translation, a focus on genre and domain continues to make sense, as the training data used by the SMT and NMT engines in operation in industry are also genre- and domain-specific, or such engines can at least be customized for these genres and domains (see Ramírez-Sánchez 2022 [this volume]).

Based on their own experience with MT, many language service providers thus recommend restricting the use of MT, and by extension, NMT, to the translation of:

Certain types of technical documentation: These usually involve already standardized texts, in which terminology use is already strict in itself, the style is direct and simple, and the use of linguistic resources closely resembles the application of controlled language rules. The conceptual framework that underpins such technical documentation may also be identical in both the source and target “locales”. The technical specifications for a personal computer that is marketed in Ireland and France remain substantially the same, for example, and so there is vast common ground when it comes to translating a text listing those specifications from English into French; extensive adaptation to take the French target user or a new conceptual framework into account is not necessary. In cases where NMT is used to translate such texts, controlled-language rules governing lexical selection or the use of pronouns like *it* probably still have some potential for application, while rules of a syntactic nature may be unnecessary.

The straightforward nature of technical specifications contrasts with marketing and legal materials associated with the same product, which might need adaptation to make them more acceptable to potential buyers, or compliant with the target legal framework. Indeed, legal translation provides one of the best examples of a domain where it is sometimes necessary to "rethink" a text completely in translation, so that it can be accommodated by a new conceptual system.

For more on the domain- and genre-specific nature of translation, see Olohan (2015) and Šarcevic (1997).

Low-risk internal documentation: These are texts which have very low visibility and where the consequences of less-than-optimal translation are not serious (see Canfora & Ottmann 2020 and Moorkens (2022 [this volume]) for more detailed discussions of risk in MT use.) They may even be limited to use within the user's or client's company. A priori, considerations such as naturalness or fluency in the target language are less relevant than would otherwise be the case (although NMT generally produces quite fluent output anyway), but companies may still wish to control lexical selection and lexical variability.

Low-risk external documentation: This refers to texts that are consulted only occasionally or sporadically, or texts that are used as a help database or similar, and that are often not produced by the client, but by the community of users of its service or product. In many such cases, the MT provider may explicitly deny liability for any losses caused by faulty translations.

MT is not usually recommended for texts of a more visible nature whose purpose is not just to inform or give instructions but also to be "appellative", that is, to arouse a particular interest in the reader, for example, in a certain brand, or to elicit a certain behaviour. In other words, the more informative a text is, the more it limits itself to the literalness of its message, the less implicit information it contains and the less it appeals to references linked to the reader's culture or social reality, the greater the expected success of MT.

4 The influence of English in controlled domains

RBMT, and later SMT, worked particularly well in environments where both language use and the nature of text genres were not figurative or creative but literal and with clear genre conventions. The source language in many cases was, without a doubt, English. The other major languages became target languages, which means that translations in certain contexts have been highly conditioned by English, giving rise to texts in which linguistic aspects have been homogenized with a view to simplifying the text to facilitate comprehension by the end reader. This is the case of genres such as user manuals for consumer goods, which, in languages such as Spanish, have—in our experience, but see also Navarro (2008) and Aixelá (2011), among others—been heavily influenced by the English source texts, both from a macro-textual point of view (in their text structure and the development of textual argumentation) and a micro-textual one (as seen in lexical, morphological and syntactic borrowings).

The objective of communication in controlled domains is to facilitate the reading and understanding of the text based on an unambiguous and precise wording, so that the original text is as easy to read as it is to translate quickly. In some industries a further step is taken and controlled languages are used to ensure that texts are free of ambiguity. In these cases, the influence of English as a source language on the other languages is much more evident.[4] The aircraft industry is an example of a context in which the rules of a controlled language are established in English and then applied in the target languages (Ghiara 2018).

These examples demonstrate the need in certain domains to control the linguistic resources used in the source text to ensure a quick and accurate translation. In these cases, the aspect of correctness of the translated text in general, and when using MT in particular, takes precedence over any other communicative aspect of the target text. Nevertheless, the arrival of NMT means that MT is now used beyond domains that are limited to specific audiences, as is described in the following section.

5 Writing for a global audience

Sometimes the objective of pre-editing is not a matter of avoiding errors in the translated text, but rather of ensuring that the translation, beyond conveying a meaning consistent with that of the source text, also achieves the same or a

[4] Seoane Vicente (2015) provides an exhaustive review of the use of English as a controlled language in different domains.

similar effect on the reader of the target text as the source text did on its reader, to the extent that this is possible. It is a question of making a text available to a global audience and attempting to have the same effect on readers in each target language.

Whether the text is to be translated with an MT system or not, from a communication perspective, for years it has been considered advisable to have the translation already in mind during the drafting phase of the source text. In fact, the preparation of documentation for translation forms part of their training for technical writers (Maylath 1997).

Over the last 50 years, the translation industry, and all related interested parties from translators to major technology developers and distributors, have learned that the best translation strategy requires appropriate internationalization of the product (Fry 2003: 14). The best way to adapt a product to any other region is to exclude those aspects which are unique to the source text region where it is being designed and developed. In this way, any digital product can be localized and used in the target language and on any device or platform, without its original design having to be modified. Something similar also appears to have happened with texts designed to be published in different languages.

Both language service providers and developers of localized digital products have found that pre-editing source texts is the key to their global communication strategy. Many language service companies advertise on their websites that good multilingual communication strategies begin with developing an appropriate source text. Digital product developers have likewise discovered that the best strategy for communication with their users and potential customers is based on keeping a global user in mind. This strategy is embodied in a set of guidelines that should be taken into account when drawing up the contents of any text. Google's documentation style guide, for example, features a basic "writing for a global audience" principle, and sets out a series of guidelines in English that facilitate the translation of documentation into any target language. These include, among others, general dos and don'ts, such as use present tense, provide context, avoid negative constructions when possible, write short sentences, use clear, precise, and unambiguous language, be consistent and inclusive (Google 2020).

Today's translation technologies make it possible to combine the use of computer-aided translation tools like translation memory tools (see Kenny 2022: §4 [this volume]) and MT systems. So, the limitations of MT in this sense are not technological, but rather determined by the quality of the raw MT output (is it error-free?) and appropriateness for the target communicative context (register, tone, genre conventions, and any other issues relevant for the translation to fulfil its communicative function).

In the case of textual genres that formally follow very rigorous conventions and essentially have an informative or instructive communicative function (for example, technical documentation, or similar), MT produced by a quality translation engine can give good or very good results, depending on the language pair and other factors. In these cases, "pre-editing" can be limited to spellchecking the source text, since these genres do not usually involve stylistic or referential features (see below) that take them outside the realm of standard and non-complex source text use.

However, genres which have a mixture of more than one communicative function, for example the recently popular "unboxing" videos for technical gadgets, which are often both instructive (informative) and entertaining (appellative and expressive), are not so simple to deal with using MT.

Texts belonging to yet other genres may contain references to the social, economic or cultural life of their source communities that allow source text readers to identify with the text, but may not have the same effect on the target text language reader (see §6.4 on referential elements). Other possible obstacles for some MT engines include rhetorical and stylistic devices (contractions, abbreviations, neologisms, incomplete sentences, etc.), that shape the source text, and with which the source text readers can identify.

NMT allows users to obtain translations with fewer and fewer errors of fluency or adequacy. It enables translations to be completed very quickly. Moreover, it seems to achieve excellent results when translating many different text genres. But a text written grammatically in the target language and without translation errors may still not be an *appropriate* translation. Pre-editing makes it possible to ensure the appropriateness of the translation with a global audience in mind. Currently, this phase is seldom used in the translation industry. In the past, some global companies using SMT or RBMT pre-edited their original texts to avoid recurring translation errors using their own systems. With NMT, pre-editing may become widespread in the industry as part of a strategy that not only avoids translation errors but also contributes to making the raw MT output appropriate to the contexts of use of the target translation.

6 Pre-editing guidelines

6.1 Opening remarks

Pre-editing is based on applying a series of specific strategies to improve MT results when preparing content for a global audience or in controlled domains. Pre-editing helps to ensure clear communication in controlled domains targeting global audiences. In this context, the predominant textual type is informational, where there is no creative or aesthetic use of language but a literal and unambiguous use with the intention of either informing or instructing the text's recipient. The following are the most common guidelines used in communication for a global audience, and are the basis for pre-editing strategies. The aim of most of these guidelines is to increase MT effectiveness in producing grammatically correct translations that reproduce the source text "message" and also to obtain translations that are appropriate to the communicative situation of the receiver according to the text function and the context in which it is used. These guidelines can be grouped into three different categories:

1. Lexical choice
2. Structure and style
3. Referential elements

Whatever the case, the success of pre-editing will be determined by two considerations. First, the function of the (source and target) text: the greater the predominance of the informative or instructive function over the phatic or aesthetic functions, the more sense it makes to pre-edit the original text. Second, the kind of errors in the raw MT output that the chosen MT system provides and that should be avoided or minimized by pre-editing the source text.

Pre-editing has two objectives: to prepare the original text so that the most error-free possible raw MT output can be obtained, and also to prepare the original text so that its translation through MT is suitable for a global audience. The pre-editing guidelines presented in this section respond to these two objectives.

6.2 Lexical guidelines

As will be seen in Pérez-Ortiz et al. (2022 [this volume]), the way each word or unit of meaning is processed in NMT is determined by its context and vice versa. A lexical choice in a text is linked to the range of texts and contexts in which the same choice is used. Let's take the case of a source text to be translated by MT and, consequently, to be published in several target languages in the shortest time possible. An appropriate choice of words in the source text can contribute not only to avoiding translation errors, but also to complying more effectively with the linguistic uses in accordance with the function of the text and the reason for its publication. Table 1 contains typical guidelines related to the lexicon.

Table 1: Typical lexical pre-editing guidelines

Guideline	Explanation
Avoid lexical shifts in register	Avoid words that can change the style of the text or the way it addresses the receiver. This facilitates understanding the text and normalizes the way the receiver is addressed.
Avoid uncommon abbreviations	Only use commonly-found abbreviations. Avoid abbreviated or reduced forms that cannot be easily translated from their immediate context.
Avoid unnecessary words	Avoid unnecessary words for transmitting the information required. Using more words than needed means that the NMT system handles more word combinations and has more opportunities to propose an inappropriate or erroneous translation.
Be consistent	Use terminology in a consistent and coherent way. Avoid introducing unnecessary word variation (that is, avoid synonymy).

6.3 Structure and style

The way a text is formulated in general, and its individual sentences in particular, are as important in terms of comprehensibility as the lexicon used. The order in which ideas are interrelated, at the sentence level, throughout a text, or even intertextually, contributes to the reader's comprehension and interpretation. In the case of NMT, the options adopted in the source text activate or inhibit translation options. An unnecessarily complex and ambiguous text structure that allows objectively different interpretations increases the possibility of the NMT system proposing correct translations of microstructural elements (terminology, phrases or syntactic units) which, when joined together in the same text, generate texts that are internally incoherent, suggest a different meaning to the source text, or are simply incomprehensible.

Table 2 gives pre-editing guidelines regarding the style and structure of the text. Most of them are not only aimed at optimizing the use of NMT systems, but also at the success of the translated text in terms of comprehensibility and meaning.

Most of the guidelines listed in Table 2 are aimed at producing a simple text that can be easily assimilated by the reader of the source text. In the case of NMT engines trained with data sets already translated under these criteria, source text pre-editing helps to obtain the best raw MT output possible. Note, however, that if an engine is trained on "in-domain" data, that is, using a specialized and homogeneous dataset, based on texts of a particular genre and related to a particular field of activity (see Ramírez-Sánchez 2022: §2.1 [this volume]), then the best possible pre-editing, if needed, will involve introducing edits that match the characteristics of that genre and domain. In addition to this general advice, in many cases it is also necessary to take into account guidelines that are specific to the source or target language. This might mean avoiding formulations that are particularly ambiguous, not only for the MT system, but also for the reader.

If we take English for instance, avoiding ambiguous expressions means, for example, avoiding invisible plurals. A noun phrase such as "the file structure" could refer to both "the structure of files" and "the structure of a particular file". Although this ambiguity is resolved as the reader moves through the text, the wording of the noun phrase itself is not clear enough to provide an unambiguous translation. Another example of ambiguous structures in many languages, not only in English, is often the way in which negation is expressed. Sentences such as "No smoking seats are available." are notorious for giving rise to different interpretations and, consequently, incorrect translations.

Verb tense forms are another aspect that may be simplified for the sake of intelligibility for the reader and error-free translations. Although the translation of the different verb tense forms and modes does not necessarily pose a problem for MT, an inappropriate use of verb tenses in the target language, despite resulting in well-formed sentences, can lead to target translation text comprehension errors. Typical guidance related to verb forms is given in Table 3.

Table 2: Aspects related to structure and style in pre-editing

Guideline	Explanation
Short and simple sentences	Avoid unnecessarily complex sentences that introduce ambiguity. This makes it easier to understand the text, both the source and translation. Syntactic structures based, for example, on anaphoric or cataphoric references may not be correctly handled by the NMT system and may lead to omissions or mistranslations. Avoid syntactic ambiguities subject to interpretation.
Complete sentences	Avoid eliding or splitting information. The compensation mechanisms for the not explicitly mentioned information typical of the source language do not necessarily work in the target language. For instance, a sentence with a verb in passive form which does not make the agent explicit can lead to misunderstanding in target texts. The same can happen when one of the sentence complements is presented as a list of options (in a bulleted list, for example). In such cases, the sentence complement is broken down into separate phrases which the NMT system may process incorrectly. Remember that MT systems normally use the sentence as a translation unit (see Kenny 2022: §7 [this volume]), i.e., the text between punctuation marks such as full stops, or paragraph breaks.
Use parallel structures in related sentences	Use the same syntactic structure in sentences in a list or that appear in the same context (e.g., section headings, direct instructions). This kind of *iconic linkage* (see Byrne 2006) usually makes it easier to understand the text, both the source and translation. In addition, it allows for the systematic identification of errors during a post-publishing phase.
Active voice	Where appropriate, use mainly the active voice or other structures that make “participants” in an action explicit (taking into account the conventions of the text genre and the languages involved).
Homogenous style	Maintain a homogeneous style. This facilitates understanding the text, both the source and translation. This is particularly related to preparing texts for a global audience.

Table 3: Aspects related to pre-editing verb forms

Guideline	Explanation
Use the active voice.	Where possible and appropriate, use the active voice.
Use simple verb tense forms; preferably the present or past simple.	Depending on your language pair and the MT engine, you may wish to avoid using compound verb forms. Although the same compound form may exist in both languages, it may not be used in the same way and may lead to different interpretations.
Avoid concatenated verbs.	Avoid unnecessary concatenations of verbs that make it difficult to understand and translate the text.

6.4 Referential elements

Referential elements are all those which substitute for or make reference to another element, whether in the same text, in the case of *intratextual* references, or outside of the text, in the case of *extratextual* references. The most illustrative example of this is pronouns, like *I*, *he*, *she*, *him*, *her*, etc., and the related category of possessive determiners, such as *my*, *his*, *her*, etc.

Pronouns with the referent in the same sentence are not usually problematic from an NMT point of view. When they are within the same sentence, there are usually no gender or number agreement problems between possessive determiners and nouns (see Bentivogli et al. (2016) for an early discussion of how the treatment of agreement phenomena in MT improved with the advent of NMT.) This is also the usually case when successive pronouns throughout the text maintain the same referent (e.g., in the case of the same subject in consecutive sentences). In other cases, however, pronouns may be translated according to the way in which the training corpus treats them most frequently. This issue is particularly sensitive when the text alternates between different referents. In these cases, even though the human reading of the text leaves no doubt as to who or what is the referent of each pronominal form, MT systems are usually unable to maintain this consistency and tend to homogenize the pronominal forms, using those that are most common.

Consequently, in the case of languages that reflect relationships between referent and pronominal forms through mechanisms such as gender and number agreement, the agreement may be lost. This particular kind of problem can usually be minimized by the use of simple sentences.

An example may help here. Example (1) contains an instance of a possessive determiner *su*, which, taken out of context, can mean either 'his' or 'her'. In example (1), the only possible interpretation is 'her', as indicated in the gloss translation provided. DeepL,[5], however, translates the instance of *su* in question as 'his', as it cannot establish the referential link between *su* and *María.*

(1) María llamó, pero Pepe no llamó. El sonido de *su* llamada me despertó.
'Maria called, but Pepe didn't call. The sound of *her* call woke me up.'
DeepL: 'Maria called, but Pepe didn't call. The sound of *his* call woke me up.'

As regards extratextual references, in addition to all the references inherent to the nature of the documentation being translated – for example, specific legislation in documents of a legal nature – there are two types of references that need to be taken into account during the pre-editing phase: 1) those that address the reader, and 2) any cultural references the reader identifies with.

As argued in the previous section, from a stylistic point of view it is best to compose simple, short sentences and use a direct style with the active voice or passive sentences that include the agent and patient. This style is especially appropriate for instructive texts. In the case of instructive texts, the direct style and active voice mean that the text always addresses a reader directly. In this type of sentence, MT tends to reflect the most frequent use found in the corpus, so if the target language allows the reader to be addressed in more than one way, it could alternate between the different options – more or less formal or explicit – and cause cohesion problems throughout the text that could be avoided by pre-editing.

Extratextual references to cultural aspects with which the reader of the source text particularly identifies are difficult to deal with generically. In many cases, pre-editing the text consists of making all the implicit information related to these cultural references as explicit as possible, keeping the global audience in mind.

In both cases (references to the reader and cultural references), MT pre-editing should take into account the target reader's language and profile.

[5] https://www.deepl.com/en/translator accessed January 2022

7 Pre-editing tools and resources

As can be seen from the previous subsections, pre-editing has to be carried out within the framework of a translation (or multilingual publication) project, so the same conditioning factors that guide a translation project need to be taken into consideration in pre-editing. A *style guide* is used to detail how to pre-edit the original texts properly. It sets out how each of the aspects should be edited in a structured manner, with examples of sentences with and without pre-editing. It is comparable to a post-editing guide, with the difference that examples are usually given in the source-text language only.

The purpose of pre-editing guides is to provide orientation about language use when preparing the text contents. These writing guidelines aim to avoid MT errors when translating the source text to different languages and to ensure that the best possible text is produced for a global audience. For this reason, they give both micro-structural (recommended or required words or recommended syntactic structures) as well as macro-structural indications on writing. The latter are aimed at providing the source text, and its subsequent translations, with the necessary mechanisms to guarantee intratextual and extratextual cohesion. When the text is embedded in a digital product or is part of the documentation linked to a consumer good, consistent use of language and referential elements (such as terminology) in all texts related to that consumer good contributes to extratextual cohesion.

A guide oriented to the preparation of contents to be published and translated also includes instructions about actions to be taken and tools that facilitate them. Comparable to the quality assurance (QA) phase in translation, preparing the text content also has to follow standards that guarantee its quality and, consequently, the success of its translation. In this case, it is not only a question of producing a grammatically correct translation, but also maintaining the client's language standard.

The actions required to control the quality of the source language content are usually as listed in Table 4.

It is important to pay attention to aspects of the text that are designed to meet with the approval of particular readerships. Inclusive language, for example, can help to avoid reader rejection of both the source and translated texts. As is the case with post-editing (see O'Brien 2022: §2 [this volume]), pre-editing attempts to avoid expressions which could be interpreted as offensive or rude. Guidelines for text content are set out by the author or publisher of the text, so for example, the pre-editing guidelines of a given company may touch upon or mandate

Table 4: QA in pre-editing

Pre-editing quality assurance (QA)	Explanation
Proofing for spelling and grammar	Guarantee that the source text is free of spelling errors which could generate comprehension difficulties by readers, or MT errors.
Using established lexicon	Check the glossary has been applied appropriately, without introducing unnecessary variation in the use of undesired synonyms. The aim here is to guarantee systematic use of terminology set out in the glossary (including trade names or proper nouns of any kind). Also, use the specialized and non-specialized lexicon as unambiguously as possible.
References to the reader	If the text is expressly addressed to the reader, then check the same style of reference is used throughout.
Style	Check that the language style is maintained consistently throughout the text. Avoid shifts in style within the same text.

gender, racial, cultural, and all kinds of inclusivity in language use. This point is particularly relevant in gender-inflected languages.

Preparing a source text for a global audience, or pre-editing, is carried out with tools that assist the writer. Most text editing programs include the most basic functions necessary to carry out pre-editing as well as QA. Other functions are available only through dedicated authoring tools. Table 5 summarizes the main functions of controlled language checkers that assist source text pre-editing.

Most editing programmes include functions that allow this type of action to be performed to one degree or another. However, when pre-editing is part of a

Table 5: Functions of controlled language checkers

Computer assisted pre-editing	Explanation
Proofing for spelling and grammar	Use the grammar and spell checker.
Using established lexicon	Use dictionaries and glossaries that establish recommended and prohibited lexical items.
References to the reader	Use the grammar checker adapted to appropriate expressions according to the pre-editing guide.
Style / register	Use the grammar checker adapted to the level of formality for the commission through suggestions concerning modifying the text.

multilingual content publishing strategy, these programmes often prove wanting. A multilingual content publishing strategy based on pre-editing and using a specific language, means homogenizing all the company or institution's communication: the content published on the web, social networks, FAQ sections, etc. In these cases, it is necessary to resort to controlled language checkers with pre-editing functions that can be integrated into the flow of content publication and, by extension, the production of translations.[6] These kinds of authoring tools usually go far beyond simply checking spelling and lexical aspects; in fact, in some cases they even provide revision proposals depending on the degree of formality of the text at any given time. In most cases, these tools are included as an additional menu in programs that are usually used to produce content, from web content managers to e-mail or social network managers. In this way, both the content author and the pre-editor can use these tools directly in the flow of creating and publishing each piece of content without the need to resort to external tools.

[6] Various controlled language checkers and other writing aids are available. Commercial tools include acrolinx (https://www.acrolinx.com/) and ProWritingAid (https://prowritingaid.com/).

8 Who should pre-edit? And when?

Texts to be translated by MT have traditionally been pre-edited in-house in large corporations as part of the technical writing phase, which requires a high level of proficiency in the source language, while knowledge of the target language(s) is not essential.

Today, the accessibility of NMT and the quality of its results allow translators in many major languages to consider including MT in their own workflows. “MT-literate” translators (Bowker & Ciro 2019) can determine whether they can benefit from incorporating NMT based on their own experience, or even by translating small samples of the source text. The typical way to assess the use of NMT is usually by analysing the resulting translated text. If the text requires limited post-editing, it is considered suitable for NMT. However, the wording of the *source text* is rarely questioned. There are texts that, due to their function and visibility, should be translatable through NMT without any major difficulties. However, if the text is unnecessarily complex, incoherent, or does not meet the established editorial standard, translating it via NMT may only accentuate these problems. In these cases, pre-editing could prove useful so that the source text has the necessary cohesion, style and use of rhetorical devices to respect the author’s intention, the function of the text and, at the same time, guarantee its comprehensibility and translatability. This is not a task which is commonly practised in the translation industry, but it is one that may become more common in the near future as a response to the need to publish content in several languages.

When the client has no previous language strategy for producing linguistic content, the skills translators have acquired make them best suited to take on the task of pre-editing texts in their mother tongue or the first foreign language into which they usually translate. Their knowledge of contrastive grammar and lexical nuances in the language combination allows them to perform the necessary edits to produce a text in the source language that is functional and understandable by the reader, and which in turn generates as few errors as possible in the text translated via NMT. In addition, their knowledge of the corresponding societies and cultures allows them to assess which referential aspects can be successful in both texts, source and target, and how to make them explicit. Their skills in both languages as well as their knowledge of both cultures and societies, mean translators are the experts best suited to prepare monolingual content intended for bilingual or multilingual publication. Their main working language will still be their mother tongue, but in this case, they do not produce a final text, but rather a machine translated text that can be understood by the end reader in another language. Thus, the post-editing phase can be minimized, although in no case should it be omitted.

Pre-editing as part of the provision of translation services makes sense whether the content is written solely to generate a translation or if it is also to be published in the source language. In both cases, pre-editing makes it possible to guarantee the quality of the original and to optimize the use of NMT.

9 Concluding remarks

The main objective of MT as a resource in translation projects is to increase productivity and, consequently, reduce the time needed to generate a good quality translation. In this sense, pre-editing manages to optimize the source text content so as to minimize errors in the translated text (when MT is used for assimilation) and the editing needed to guarantee the expected quality (when MT is followed by post-editing or used as a resource for human translation).

When NMT is capable of producing translations with virtually no fluency or adequacy errors in informative or instructive texts, then the challenges for MT go beyond these text types. However, translating texts with different communicative functions, such as for games or texts of a more appellative nature, is not only a matter of avoiding errors. It is necessary to produce a translation that is in line with the intention of the source text and with which the target reader can identify in the same way as the source text reader. In this case, pre-editing takes on an added value: the preparation of a text suitable for publishing multilingual content.

As a strategy, pre-editing may play a certain role in foreign language learning. But its main environment is in multilingual content publishing. Although it was originally part of translation workflows for technical documentation and the like, the expansion of NMT could lead to pre-editing being applied to texts of a more complex nature, or even to translators eventually putting their skills at the service of the source text, instead of focusing on the target text, as has happened throughout centuries of translation history.

References

Aixelá, Franco. 2011. An overview of interference in scientific and technical translation. *The Journal of Specialised Translation* 11. 75–88. https://www.jostrans.org/issue11/art_aixela.pdf.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo & Marcello Federico. 2016. Neural versus Phrase-Based Machine Translation quality: A case study. In *EMNLP 2016*. arXiv:1608.04631v1.

Bowker, Lynne & Jairo Buitrago Ciro. 2019. *Machine translation and global research*. Bingley: Emerald Publishing.

Byrne, Jody. 2006. *Technical translation. Usability strategies for translating technical documentation*. Dordrecht: Springer.

Canfora, Carmen & Angelika Ottmann. 2020. Risks in neural machine translation. *Translation Spaces* 9(1). 58–77.

Castilho, Sheila, Stephen Doherty, Federico Gaspari & Joss Moorkens. 2018. Approaches to human and machine translation quality assessment. In Federico Gaspari Joss Moorkens Sheila Castilho & Stephen Doherty (eds.), *Translation quality assessment: From principles to practice*, 9–38. Cham: Springer.

Drugan, Joanna. 2013. *Quality in professional translation: Assessment and improvement*. London: Bloomsbury.

Fry, Deborah. 2003. *The localization industry primer*. 2nd edition. Updated by Arle Lommel. Féchy: LISA. https://www.immagic.com/eLibrary/ARCHIVES/GENERAL/LISA/L030625P.pdf.

Gerlach, Johanna. 2015. *Improving statistical machine translation of informal language. A rule-based pre-editing approach for french forums*. Doctoral thesis. University of Geneva. https://archive-ouverte.unige.ch/unige:73226.

Ghiara, Silvia. 2018. *El lenguaje controlado. La eficacia y el ahorro de las palabras sencillas*. https://qabiria.com/es/recursos/blog/lenguaje-controlado.

Google. 2020. *Writing for a global audience. Google developer documentation style guide*. https://developers.google.com/style/translation.

Hiraoka, Yusuke & Masaru Yamada. 2019. Pre-editing plus neural machine translation for subtitling: effective pre-editing rules for subtitling of TED talks. In *Proceedings of machine translation summit XVII: translator, project and user tracks*, 64–72. Dublin: European Association for Machine Translation. https://aclanthology.org/W19-6710.

Kenny, Dorothy. 2022. Human and machine translation. In Dorothy Kenny (ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence*, 23–49. Berlin: Language Science Press. DOI: 10.5281/zenodo.6759976.

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

Maylath, Bruce. 1997. Writing globally: Teaching the technical writing student to prepare documents for translation. *Journal of Business and Technical Communication* 11(3). 339–352.

Miyata, Rei & Atsushi Fujita. 2017. Dissecting human pre-editing toward better use of off-the-shelf machine translation systems. In *Proceedings of the 20th annual conference of the european association for machine translation (EAMT)*, 54–59. https://ufal.mff.cuni.cz/eamt2017/user-project-product-papers/papers/user/EAMT2017_paper_42.pdf.

Miyata, Rei & Atsushi Fujita. 2021. Understanding pre-editing for black-box neural machine translation. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics*, 1539–1550. https://aclanthology.org/2021.eacl-main.132.pdf.

Moorkens, Joss. 2022. Ethics and machine translation. In Dorothy Kenny (ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence*, 121–140. Berlin: Language Science Press. DOI: 10.5281/zenodo.6759984.

Navarro, Fernando A. 2008. La anglización del español: Mucho más allá de bypass, piercing, test, airbag, container y spa. In Luis González & Pollux Hernúñez (eds.), *Traducción: Contacto y contagio. Actas del III congreso internacional « el español, lengua de traducción ». 12-14 July 2006*, 213–132. Puebla: ESLEtRA. https://cvc.cervantes.es/lengua/esletra/pdf/03/017_navarro.pdf.

O'Brien, Sharon. 2003. Controlling controlled English. An analysis of several controlled language rule sets. In *Controlled language translation*. Dublin City University. 15-17 May 2003. EAMT/CLAW. https://aclanthology.org/2003.eamt-1.12.pdf.

O'Brien, Sharon. 2022. How to deal with errors in machine translation: Post-editing. In Dorothy Kenny (ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence*, 105–120. Berlin: Language Science Press. DOI: 10.5281/zenodo.6759982.

Olohan, Maeve. 2015. *Scientific and technical translation*. London: Routledge.

Pérez-Ortiz, Juan Antonio, Mikel L. Forcada & Felipe Sánchez-Martínez. 2022. How neural machine translation works. In Dorothy Kenny (ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence*, 141–164. Berlin: Language Science Press. DOI: 10.5281/zenodo.6760020.

Ramírez-Sánchez, Gema. 2022. Custom machine translation. In Dorothy Kenny (ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence*, 165–186. Berlin: Language Science Press. DOI: 10.5281/zenodo.6760022.

Rossi, Caroline & Alice Carré. 2022. How to choose a suitable neural machine translation solution: Evaluation of MT quality. In Dorothy Kenny (ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence*, 51–79. Berlin: Language Science Press. DOI: 10.5281/zenodo.6759978.

Šarcevic, Susan. 1997. *New approach to legal translation*. The Hague: Kluwer Law International.

Seoane Vicente, Ángel Luis. 2015. *Lenguaje controlado aplicado a la traducción automática de prospectos farmacéuticos*. handle.net/10045/53587. Doctoral Thesis. URI: http://hdl.

Seretan, Violeta, Pierrette Bouillon & Johanna Gerlach. 2014. A large-scale evaluation of pre-editing strategies for improving user-generated content translation. In *Proceedings of the 9th edition of the language resources and evaluation conference (LREC)*, 1793–1799. http://www.lrec-conf.org/proceedings/lrec2014/pdf/676_Paper.pdf.