Chapter 3

How to choose a suitable neural machine translation solution: Evaluation of MT quality

Caroline Rossi
Université Grenoble-Alpes

Alice Carré
Université Grenoble-Alpes

Machine translation (MT) is evolving fast, and there is no one-size-fits-all solution. In order to choose the right solution for a given project, users need to compare and assess different possibilities. This is never easy, especially with MT outputs that look increasingly good, thus making mistakes harder to spot. How can we best define and assess the quality of a neural MT solution, so as to make the right choices? The first step is certainly to define needs as precisely as possible. Having defined a pragmatic view of quality, we introduce the key notions in human and automatic evaluation of MT quality and outline how they can be applied by translators.

1 Introduction

Beyond the hype about neural machine translation (NMT), users do notice that machine-translated texts have been getting better. The main point of this chapter is to show that even though machine translation (MT) outputs may appear to be more fluent than before, they are not necessarily easier to deal with. Besides, NMT outputs are likely to vary, and should be considered in context and according to the needs of end-users. In what follows, we suggest definitions of quality

Caroline Rossi & 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

and measures that can be used to reach beyond the apparent ease and fluency of NMT outputs.

The overarching question that this chapter seeks to answer is: how can NMT solutions be assessed in a trustworthy and useful way? The answer may vary, for example, according to use cases and text types. In what follows, we explain the key issues with MT evaluation, with a view to helping users to choose an MT engine that suits their specific needs.

1.1 Assessing Machine Translation quality: What are we talking about?

Starting from a broad definition of quality in translation, we see that it covers both product and process: "Quality in translation is both the quality of an end-product (the translated material) and the quality of the transaction (the service provided)" (Gouadec 2010: 270). Besides, translation quality assessment will very much depend on the context in which the translation is done, and the expectations of translator trainers will certainly differ from those of a client who needs a translation for specific purposes. In other words, "the notion of quality is relative" (Grbić 2008: 232). In translation studies, quality has been notoriously difficult to define and inevitably variable: a number of review studies thus relate changes in translation theories to changing views of quality (see e.g. Drugan 2013; House 2015). And when it comes to assessing the quality of MT outputs, different definitions of quality are again used. "The evolution and widespread adoption of translation technologies, especially machine translation (MT), have resulted in a plethora of typically implicit and differently operationalized definitions of quality and respective measures thereof" (Doherty 2017: 131). As far as MT is concerned, quality has been seen more as a means to an end (namely improving systems), and so a pragmatic approach has prevailed, often involving a combination of human and automatic evaluation (Doherty 2017: 133). Before moving on to present existing means of evaluation and how they might be combined, let us explain why a pragmatic approach is needed.

From the point of view of MT users, the quality of an MT output is a complex thing to assess. Indeed, if quality crucially depends upon the system used, the context of the translation and the needs of end-users are also key factors to take into account. Consider a fairly simple example: you have probably found it easy to adjust to poorly translated instructions in a user manual, because you already had quite a clear idea of how to use the product you had bought, or because the pictures were enough to guide you. In such circumstances, we are likely to

find relatively high tolerance for MT errors (Castilho & O'Brien 2016). Now consider a completely different setting that also involves technical texts, but with an added legal dimension: users of translated patents need precise and relevant information, so tolerance for MT errors will be much lower. Looking at NMT for the patent domain, Castilho et al. (2017: 113) have, for instance, evidenced a tendency of NMT to omit elements from the source text, in a context where a piece of information missing from the machine-translated text may have serious consequences. In both cases, a pragmatic approach to quality assessment would imply using measurable indicators of usefulness, such as user satisfaction ratings, productivity increases in post-editing, or increased sales based on machine-translated descriptions of products.

Overall, assessing translation quality is far from trivial, and several factors come into play when evaluating a translation, whether it is done by humans or machines. For a start, there is usually more than one valid solution in translation: the same source text can have several translations, all equally acceptable. What is more, if the evaluation of a translation is entrusted to human evaluators, the evaluation process will often be subjective: indeed, it is not uncommon for evaluators to disagree on the level of quality of a given translation. Evaluations based on what humans *do* with translations can be objective, however, when they use productivity measures. Overall, in order to compensate for subjectivity, it is essential to clearly define the objectives and indicators of each evaluation. Another disadvantage of human evaluation is that it is also a time-consuming and resource-intensive process. As an alternative to human evaluation, it is possible to use algorithms to carry out an automatic evaluation, which is certainly cheaper and faster than human evaluation, but also sometimes less relevant, because it may not track usefulness in a particular application. Both types of evaluation thus have advantages and disadvantages; and your choice should depend above all on your translation project and needs.

1.2 Good-enough quality: Think twice!

Using a pragmatic approach to MT assessment, as proposed above, it becomes clear that not all MT errors or approximations have the same impact. Cooking recipes often provide us with a good testing ground, because the results of the translation are easy to see (and taste), and as a matter of fact they have been used for some time now to produce jokes about MT errors. Here is an example of a lasagne recipe, machine translated from French into English, together with a question: would you be able to make the lasagne if you could use only this MT output?

Table 1: Examples of NMT errors (underlined) found in a translated recipe (FR>EN)

French source text[a]	English NMT output
Préchauffez votre four à 180°C. Dans un plat à gratin beurré, versez un peu de béchamel. Déposez une couche de pâte et poser une couche de farce a la viande. Déposez à nouveau des pâtes, béchamel et viande.	Preheat your oven to 180°C. In a buttered gratin dish, pour a little béchamel. Put a layer of dough and a layer of meat stuffing. Put a layer of pasta, béchamel and meat filling on top.
Terminer par une couche de pâte avec de la béchamel et saupoudrez de fromage râpé. Laissez cuire 35 à 40 minutes.	Finish with a layer of dough with béchamel and sprinkle with grated cheese. Leave to cook for 35 to 40 minutes.

[a]Lasagnes de grand-mère (French recipe): https://www.750g.com/lasagnes-r66998.htm

You would certainly be surprised at the result, since variation in the French source text between the singular and plural of "pâte(s)" results in the appearance of dough in a recipe that really only includes pasta. You might be cautious and cunning enough to guess, but trusting the rather fluent MT output would have resulted in baking a different dish, and the mistake was induced by just one letter (a plural ending). Even though the machine-translated text is very fluent and reasonably accurate on the whole, and would require only small changes to improve it, we see that one serious issue is enough to make the translation dysfunctional. The recipe's relative simplicity, together with knowledge about a common dish, could help readers work their way around this problem, but with many other text types and specialized domains these elements won't apply. What is more, the evaluation method proposed here involves humans, ingredients and a kitchen: it would be a very expensive test, and one that is hardly ever used.

Besides such misfires, most MT users are likely to encounter problems with abstract notions and metaphorical expressions. In the example in Table 2, which shows part of the blurb for a book published in French alongside its translation into English by an NMT system, would an English-speaking reader be able to guess that "a veritable pie in the sky" (English MT output for the French "véritable tarte à la crème") meant a well-trodden path or prefabricated subject?

If you're already used to dealing with MT, you probably recognise these mistakes, and a number of others. Experience makes a difference! And the more

Table 2: Example of mistranslation of an idiomatic expression (underlined)

French source text[a]	English NMT output
L'indépendance du parquet, véritable tarte à la crème remise sur le plateau à chaque campagne présidentielle, était aussi une proposition du candidat Macron. Il ne l'a pas tenue.	The independence of the public prosecutor's office, a veritable pie in the sky put forward in every presidential campaign, was also a proposal by candidate Macron. He did not keep it.

[a]French source text: https://www.grasset.fr/livres/ministere-de-linjustice-9782246827504

fluent the MT output, the more caution is needed: recent studies have shown that students' correction rates were lower with NMT than other less fluent types of MT (Yamada 2019). Getting used to the recurrent problems found in NMT outputs for a given language pair (and domain) will help you detect them and fix them more efficiently.

To conclude, even though NMT quality has undeniably been getting better, it is probably not easier to deal with than other types of MT, and MT is never a simple recipe for success. Instead, you will need to pay attention to small mistakes hidden in a fluent MT output, and to carefully consider your needs before deciding whether an MT solution is appropriate.

2 Choosing a suitable MT engine for specialized translation purposes

While the previous section has been mostly concerned with general aspects of NMT use and assessment, choosing a suitable MT engine for professional translation purposes means taking into account a variety of other aspects. Crucially, an MT engine that suits your professional purposes has to respect the privacy requirements of your client, be seamlessly buildable into your workflow, offer your language pairs and provide you with an output that you can post-edit with minimal effort to meet your client's needs. The quality of the output will depend on the specialized domain and text type, the trainability of the engine, and the pre-editing and post-editing effort you are willing to put in.

2.1 Privacy and confidentiality

How does an MT system handle your data? Even though this question is a vital one for you and your commissioner or client, most MT solutions do not provide you with a clear answer or any kind of warning at first sight. Instead, you will have to read privacy statements carefully and make sure you have made the right choice before you start.

For instance, this warning is found in the privacy statement of eTranslation, the EU's NMT system: "Users should exercise their judgement when submitting potentially sensitive documents to any online service, including eTranslation". Such warnings are valid even for solutions in which the data is not kept on the provider's servers, and privacy issues may indeed arise from simple situations in which data has been transferred only to be deleted a few moments later. Concerns about confidentiality mean that the use of even internal MT systems like eTranslation has been forbidden in hearings at the Court of Justice of the European Union. This is because no MT system could possibly conform to the strict confidentiality requirements for such hearings[1] (C. Lenglet, personal communication).

Of course, the risk that privacy issues will arise with free online MT is considerably higher, because data is kept and reused constantly. Thus, some of the confidential segments of a text inadvertently fed to a free MT system could be leaked in unexpected ways (for more information on how to handle your data ethically and safely, see Moorkens 2022 [this volume]).

2.2 Comparing MT outputs

You may wish to test several MT tools with the same source text and compare the outputs in order to identify the best tool for your needs.

There are several ways to compare outputs depending on the question you seek to answer. The scores used for comparison may, for example, be based on human assessment, automatic evaluation and or the measurement of post-editing (hereafter PE) effort. The measurement of PE effort is covered in more detail in O'Brien (2022 [this volume]). It suffices to say here that an MT output that requires little or no PE is considered "better" than one that requires a lot of PE, and that PE effort is used to measure quality in cases where there is a realistic assumption that machine translated texts will be used for dissemination purposes. (See Kenny (2022 [this volume]) on the distinction between MT for assimilation and MT for dissemination.) In the rest of this chapter, we focus on human assessment and automatic evaluation of MT outputs.

[1] As described and explained here: https://eur-lex.europa.eu/legal-content/en/TXT/PDF/?uri=CELEX:32013D0488

2.3 Human evaluation

Human assessment relies on human evaluators assessing the output of one or more systems. This is usually done sentence by sentence (or "segment" by "segment"), but document-level evaluations have also been carried out (see, for example, Castilho 2020). Human evaluators are usually asked to score each segment using two different criteria. The first, known as *adequacy*, measures the amount of meaning in the source segment that is rendered in the machine translated segment. Adequacy is usually measured on an ordinal scale: a typical scale ranges from 1 (understood as indicating that none of the meaning expressed in the source segment is expressed in the machine translated segment) to 4 (understood as meaning that all the meaning expressed in the source segment is expressed in the machine translated segment). Sometimes five-point scales are used, but with an uneven number of points in the scale, there is always a risk that evaluators will overuse the middle point. The second criterion, known as *fluency*, measures "the extent to which the translation follows the rules and norms of the target language" (Castilho et al. 2018: 18). In principle, fluency judgements can be made without the evaluator even looking at the source segment. They also usually rely on four-point ordinal scales, with 1 indicating that there is "no fluency" in the machine translated segment, while 4 denotes a native-like segment (Castilho 2020: 1152-1153). Adequacy and fluency evaluations are generally considered to be extremely time-consuming and, therefore, expensive to conduct.

A more straightforward — and thus faster — comparison of outputs from two different MT systems can be conducted by simply asking evaluators to rank the outputs, that is, to say which one is "better" without specifying why. This approach has been used by many MT providers to get fast feedback from online users. A good example was Microsoft's use of this approach to get user evaluations of outputs from its statistical and neural MT systems in 2017, as reported in Moorkens (2018).

Other MT providers have developed more elaborate interfaces to assist in human evaluations of MT outputs. Kantan AI, for example, offers a tool call KantanLQR (for "Language Quality Review") which allows users to specify which quality criteria (for example, adequacy, fluency, terminology use, etc.) are most important for their purposes and then to compare up to four different MT outputs, based these quality indicators.[2] Tools like this are particularly useful, as they provide visualizations, often in the form of pie charts and bar charts, of human evaluators' scores for individual segments, and they can compute overall

[2]For more information, see: https://kantanmt.zendesk.com/hc/en-us/articles/115003644483-What-is-KantanLQR- and https://twitter.com/i/status/1466392446552657927.

scores for different MT engines or systems. They also typically have functions designed for use by project managers in translation companies, as well as the actual evaluators. Non-commercial tools such as PET (Aziz et al. 2012) are also available to help in human evaluations of MT outputs, and are frequently used by academic researchers.

Other familiar tools that can be used to support human evaluations include spreadsheet programs. These allow manual input of scores into tables like that suggested in Table 3. In-built functions can then be used to compute average scores for the quality indicators you have used. A variety of free-to-use online forms can also be used to conduct human evaluations.[3] These are particularly useful for conducting surveys, and can often automatically compute summary and other statistics in the same way that spreadsheets do.

Table 3: Suggested spreadsheet for comparing MT solutions

Source text	NMT1	NMT2	NMT3	Preferred output	Comments
Segment 1					
Segment 2					

2.4 Error typologies

Sometimes human evaluators are asked not just to score a machine translated segment or document using one of the metrics described above, but to say precisely what is wrong with the particular output, by assigning each error in the segment to a category specified in an *error typology*. Categorizing errors is an important step in diagnosing problems in MT output, often in an effort to provide feedback to system developers.

Error typologies tend to be rather complex, however. The Multidimensional Quality Metrics (MQM) framework, for instance, includes an extensive list of error categories (Mariana et al. 2015: 140). For the sake of simplicity and ease, a limited set of common errors could be used in an evaluation, such as the one selected by Moorkens for a practical in-class translation evaluation exercise (Moorkens 2018: 380). It includes:

- Word order errors (incorrect word order at phrase or word level)

[3]Perhaps the best known example is Google Forms. See https://support.google.com/docs/answer/6281888?hl=en&co=GENIE.Platform%3DDesktop

- Mistranslations (incorrectly translated word, wrong gender, number, or case)
- Omissions (words from the source text have been omitted from the target text)
- Additions (words not in the source text have been added in the target text)

It might turn out that the small sample of MT output selected for evaluation is not representative enough of each engine's performance, and, ideally, the comparison should be repeated on different samples before choosing the best engine or system. However, while large institutions may have the means of conducting large-scale evaluation campaigns, smaller translation services and freelancers may do better to turn to automatic metrics and measuring post-editing effort.

2.5 Using automatic evaluation metrics

Being much faster and cheaper than human assessment, automatic evaluation metrics (AEMs) allow MT users to assess the quality of MT outputs as frequently as required. For example, if you are training your engine, running tests after each change allows you to check whether your engine has improved for your purposes. If you are not training an engine yourself, but are able to choose one from among several for a given project, AEMs allow you to evaluate multiple MT outputs for the same source text sample.

Human translators often produce widely different translations for the same source text. MT systems therefore cannot be expected to match a human translation exactly. But because a machine translation that is very similar to a human translation might be better than one that differs greatly from it, many AEMs are based on the principle of similarity: the evaluation tool is fed both a human-generated "gold standard" or *reference translation*, and the system output, known as the *candidate translation* (sometimes called *hypothesis*). It then compares the candidate against the reference translation and computes the similarity or dissimilarity. To take variation across reference translations into account, some evaluation tools can be fed multiple reference translations.[4]

A large number of AEMs, or variations of existing AEMs, have been proposed over the last two decades. In this section we concentrate on just a handful of AEMs however, basing our selection on what readers are likely to encounter in

[4] In such cases, a decision needs to be made on how to compute the length of the reference translation. Multi-reference BLEU, for example, uses the length of the reference "closest in size to the candidate translation" (Qin & Specia 2015: 114)

the evaluation interfaces mentioned elsewhere in this chapter, namely KantanMT and MutNMT. Readers interested in expanding their knowledge of AEMs beyond those covered here are referred to Koehn (2010) and Koehn (2010). In the interests of consistency, we follow the terminology and notation used by Koehn and apply his explanations, where applicable, in the evaluation of NMT outputs in a worked example below.

The example is based on an excerpt from a user manual for a transceiver,[5] and is reproduced in Figure 1.

Source text:	Battery pack is attached to the transceiver.

Figure 1: source text of the main example for this section

The sentence appears in a bulleted list of conditions under which the transceiver is guaranteed to be water-resistant (see Figure 2).

Important note
Water resistance of the transceiver (IP57: 1 meter / 30 minutes) is assured only when the following conditions (*sic*):

- Battery pack is attached to the transceiver;
- Antenna is connected to the antenna jack;
- and MIC/SP cap is installed in the MIC/SP jack.

Figure 2: excerpt from a transceiver user manual

While it is written for the general public, the source text is a technical text and its translation would thus constitute a specialized translation task. It addresses the domain of radio communication, and therefore has to respect the terminology and phraseology of that domain, and its genre is that of a user manual, which means in turn that it should follow the conventions of such documents. For example, each concept should be referred to using one term only (i.e., synonyms are not permitted), and each term should correspond to one concept only (a property known as *monosemy*), instructions should be kept short and simple, and instructions should all be written following the same pattern. (For more on domains and genres, see Kenny 2022 [this volume].) In our proposed translation project, the

[5]The VX-450 series of Vertex Standard, now discontinued.

target text will have to be translated into French and will have the same function as the source text: it will be provided to customers along with the transceiver.

In what follows, we will consider this excerpt and the way it was translated by three different MT tools. The first one (hereafter *system A*, the output of which is called *candidate A*) is eTranslation, the EU's MT tool.[6] The second one (hereafter *system B*, which outputs *candidate B*) is Google Translate.[7] The third one (hereafter *system D*, which outputs *candidate D*) is DeepL Translator.[8] At the time of writing, these systems are freely accessible to the general public, with one proviso: eTranslation requires would-be users to register and to belong to one of three categories of users: SMEs, Public Service Officials and Public Sector Service Providers.

Some of the more basic AEMs that we will present in this section can be computed by hand. However, for the more complex metrics we will use MutNMT to compute scores.[9] While we are presenting an example to give readers an idea of how these metrics work, we would like to make it clear here that the exact computation of an AEM score varies depending on the particular implementation details of each metric: if you use different tools to compute what seems to be the same AEM (say, BLEU, for instance), you may well get different results.[10] The discrepancy in results may have its origins in the way the tool deals with quotation marks, hyphens, breaking and non-breaking spaces, etc., before computation, in the way it defines tokens (does it take into account apostrophes, hyphens, punctuation or linguistic information such as lemmas or multiple-word units?), in its sensitiveness to case, or in metric parametrization specifics (e.g., what order of *n*-grams is used for the exact implementation?).[11] In our example, we changed the apostrophes in candidate D to those used in the reference translation. This way, the different coding of smart and straight quotes will not interfere with the AEM results, and we can focus on the translation output *per se*. Furthermore, when we compute AEMs by hand for the purposes of explanation, we consider hyphens and apostrophes as word "breaks". This means that the total word count of our reference translation (see Figure 3) is eight.

[6] https://webgate.ec.europa.eu/etranslation/translateTextSnippet.html

[7] https://translate.google.com/?hl=en

[8] https://www.deepl.com/en/translator. Note that we are referring to DeepL as "system D" and not "system C" in order to avoid confusion in cases where we use *C* to refer to a candidate translation.

[9] https://mutnmt.prompsit.com/index

[10] Note, however, that there has recently been an effort to normalize and group reference implementations of AEMs in software such as Matt Post's sacrebleu (Post, 2018).

[11] The authors would like to thank Gema Ramírez-Sánchez for her explanations.

Reference translation:	La batterie est installée sur l'émetteur-récepteur.

Figure 3: reference translation for example

Figure 4 shows the source text, candidate translations and reference translation that we will consider in what follows.

Source text:	Battery pack is attached to the transceiver.
Reference:	La batterie est installée sur l'émetteur-récepteur.
Candidate A (eTranslation):	Le bloc-batterie est fixé à l'émetteur-récepteur.
Candidate B (Google Translate):	La batterie est fixée à l'émetteur-récepteur.
Candidate D (DeepL) :	Le bloc-piles est fixé à l'émetteur-récepteur.

Figure 4: Main example for this section – source text, reference translation and candidate translations

What can a human evaluator say about these examples? Firstly, the term "battery pack" should be translated by "batterie", unless the customer has specified otherwise. The translation "bloc-piles" (candidate D) is plainly wrong: this transceiver does not function on "piles", which are electrochemical cells designed to be used once and then discarded, but rather on a "batterie", that is a rechargeable pack of cells. In this case, the transceiver operates on a lithium-ion battery. Talking of "bloc-batterie" (candidate A) is not intrinsically incorrect. Rather, it is not idiomatic; it is a calque, i.e. an overly word-for-word translation, of the English sentence. Secondly, the verbal form "is attached to" can just as well be rendered by "est installée sur" or by "est fixée à": this is a matter of personal preference. Thirdly, "transceiver", which is a contraction of "transmitter-receiver" should ideally be translated by "émetteur-récepteur", as is the case in all translations shown in Figure 4. However, "radio" or even "appareil" ("device") would have worked just as well for the purposes of this translation project (for more on translation and equivalence, see Kenny 2022 [this volume]). Now, let us take an in-depth look at how AEMs would assess these candidate translations.

2.5.1 Core concepts: *n*-grams, precision, recall and F-measure

In this section, we present four concepts that constitute the building blocks of the more elaborate AEMs we present in subsequent sections: *n-grams*, *precision*, *recall* and *F-measure*.

2.5.1.1 *n*-grams

n-grams (see Kenny 2022 [this volume]) are normally understood in translation as *n*-word sequences. In our example sentence, "battery" is a 1-gram or *unigram*, "battery pack" is a 2-gram or *bigram* and "battery pack is" is a 3-gram or *trigram*. Other orders of *n*-gram are simply called 4-gram, 5-gram, etc., making "battery pack is attached" a 4-gram.

n-grams are commonly used in language modelling, where, for example, a trigram probability states the probability of seeing a word given that you have already seen the two words before it.

When we discuss AEMs, *n-grams* are merely *n*-word sequences in the candidate translation that also occur in the reference translation. More recently, AEMs have been proposed which consider sequences of characters instead of words. *N*-grams are then understood as sequences of *n* characters, rather than sequences of *n* words.

We will be using the notion of *n-grams* as *n-word sequences* when discussing BLEU (see 2.5.4.), and of *n-grams* as *n-character sequences* when discussing ChrF3 (see 2.5.5).

2.5.1.2 Precision and recall

Precision is a very basic concept used in various branches of natural language processing. It can be explained using a very simple example. If a teacher asked a student to name the days of the week in English, and the student replied "Monday, Tuesday", then the student would have given two correct answers and no incorrect answers. Because precision is understood as the ratio of correct answers given to the total number of answers given, the student would score two out of two, which is equal to an impressive precision score of 100%.

But the teacher would see the student's answer as very problematic, because the teacher knows what the answer *should have been*, and that the student has neglected to mention five of the seven days of the week. The teacher could therefore object that the student's *recall* is bad, where recall is understood as the ratio of correct answers given to the total number of correct answers (in the ideal reply). In this case, the student's recall score would be two out of seven, or just under 29%.

In the context of the automatic evaluation of MT, precision computes the ratio of correct words in the candidate translation, i.e., those that also occur in the reference translation, to the total number of words in the candidate:

$$\text{precision of } C = \frac{\text{no. of correct words in } C}{\text{no. of words in } C} \quad (1)$$

where C is the candidate translation.

Let us consider our example. In Figure 5, the candidates are compared against the reference, and "correct" words, i.e. words that also occur in the reference, are underlined, while "incorrect" words, i.e. words that do not occur in the reference, are crossed out.

Source text:	Battery pack is attached to the transceiver.
Candidate A (eTranslation):	~~Le bloc-~~batterie est ~~fixé à~~ l'émetteur-récepteur.
Candidate B (Google Translate):	La batterie est ~~fixée à~~ l'émetteur-récepteur.
Candidate D (DeepL):	~~Le bloc-piles~~ est ~~fixé à~~ l'émetteur-récepteur.
Reference:	La batterie est installée sur l'émetteur-récepteur.

Figure 5: source text, reference translation, candidate translation

Let us now work out the precision of each candidate. System A's output has five correct words out of a total of nine, which gives a precision of 0.56, or 56%.[12] System B's output has six correct words out of eight, so its precision score is 0.75, or 75%. Finally, system D's output has four correct words out of nine, which gives a precision of 0.44, or 44%. According to this metric, system B's output is better than that of system A or system D.

Recall, in the same context, computes the ratio of correct words in the candidate to the total number of words in the *reference*:

$$\text{recall of } C = \frac{\text{no. of correct words in } C}{\text{no. of words in } R} \quad (2)$$

where C is the candidate translation, and R the reference translation.

In other words, recall takes into account not just what the candidate translation said, but what it should have said.

Let us go back to our example (Figure 6).

The reference translation comprises eight words. System A's output has five correct words out of a total of eight in the reference translation, and thus a recall ratio of 0.63, or 63%. System B's output has a total of six correct words, which makes for a recall of 0.75, or 75%, while system D's output has a total of four correct words, which makes for a recall of 0.5, or 50%. According to this metric, system B's output is again better than system A's or D's.

[12] In this section, all results will be rounded to the nearest hundredth when dealing with score ranges comprised between 0 and 1, and to the nearest unit when dealing with percentages.

Source text:	Battery pack is attached to the transceiver.
Candidate A (eTranslation):	~~Le bloc~~-batterie est ~~fixé à~~ l'émetteur-récepteur.
Candidate B (Google Translate):	La batterie est ~~fixée à~~ l'émetteur-récepteur.
Candidate D (DeepL):	~~Le bloc-piles~~ est ~~fixé à~~ l'émetteur-récepteur.
Reference:	La batterie est installée sur l'émetteur-récepteur.

Figure 6: Source text, reference translation, candidate translation

2.5.1.3 *F*-measure

The student in our example above could choose to prioritize precision over recall and thus refuse to give any more answers after "Monday, Tuesday", because they do not want to risk giving a wrong answer. Alternatively, they might choose to prioritize recall by blurting out tens of answers in the hope that enough of them are actually correct. They thus might reply "Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday, January, February, March, April, May, June, July, August, September, October, November, December". Their recall would now shoot up to 100% as they would have given seven out of seven correct answers (for the days of the week), but their precision will plummet to under 37% as only seven out of the nineteen answers in their reply are correct. From the teacher's point of view, neither strategy is ideal. What the teacher wants is for the student to optimize both precision and recall as the same time. They need a score that combines both. This is where the F-measure comes in.

In mathematical terms, the *F-measure* is the harmonic mean of precision and recall. It is computed as follows:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

which can also be reformulated as:

$$F = 2 \cdot \frac{\text{no. of correct words in } C}{\text{no. of words in } C + \text{no. of words in } R} \tag{4}$$

where C is the candidate translation, and R the reference translation.

Let us compute the F-measure of our three candidate translations in Table 4.

System A's output gets an F-measure of 59%, while system B's output gets an F-measure of 75% and system D's output an F-measure of 47%. According to this metric, system B's output is still better than system A's or D's.

Table 4: Precision, recall and F-measure of candidates A, B and C

Metric	Candidate A	Candidate B	Candidate D
precision	56%	75%	44%
recall	63%	75%	50%
F	$2 \cdot \frac{56 \cdot 63}{56+63} = 59$	$2 \cdot \frac{75 \cdot 75}{75+75} = 75$	$2 \cdot \frac{44 \cdot 50}{44+50} = 47$

With the three metrics *precision*, *recall* and *F*, the higher the score, the better the MT output is deemed to be. However, these metrics work at the word level and do not take word order into account.

2.5.2 Translation error rate (TER)

The *translation error rate* (also called translation *edit* rate, or TER) takes word order into account.

It is based on the *word error rate* (WER), which uses the Levenshtein distance. The Levenshtein distance computes the difference between sequences (in the present case, sequences of words); it is defined as "the minimum number of editing steps – insertions, deletions, and substitutions – needed to match two sequences" (Koehn 2010: 224). WER then normalizes that distance by the length of the reference (Koehn 2010: 225):

$$\text{WER} = \frac{\text{no. of substitutions} + \text{no. of insertions} + \text{no. of deletions}}{\text{no. of words in } R} \tag{5}$$

where R is the reference translation.

However, when sequences of words or indeed whole clauses are moved elsewhere in a sentence, each word move counts as two errors (one deletion and one insertion), which can result in a very poor WER.

TER solves this by adding a shift operation, which means that moving any sequence of words counts only as one error:

$$\text{TER} = \frac{\text{no. of shifts} + \text{no. of substitutions} + \text{no. of insertions} + \text{no. of deletions}}{\text{no. of words in } R} \tag{6}$$

where R is the reference translation.

Let us go back to our example. Compare candidate translations A, B and D with the reference translation (Figure 7). What is the minimum number of steps required to go from candidates A, B and D to the reference translation?

Source text:	Battery pack is attached to the transceiver.
Candidate A (eTranslation):	Le bloc-batterie est fixé à l'émetteur-récepteur.
Candidate B (Google Translate):	La batterie est fixée à l'émetteur-récepteur.
Candidate D (DeepL):	Le bloc-piles est fixé à l'émetteur-récepteur.
Reference:	La batterie est installée sur l'émetteur-récepteur.

Figure 7: Source text, candidate translations A, B and C, and reference translation

TER is a heuristic process, i.e. an iterative process, where the algorithm tries to find the best solution (the minimal number of steps required to go from one sequence to another) by testing successive hypotheses. To calculate the TER manually, one can use a matrix. However, we are going to propose a shorter, if imperfect way,[13] for the sake of explanation: let us compare each candidate translation, and count the number of matches, shifts, substitutions, additions and deletions. Remember, the number of matches will not go into the final calculus. As we mentioned before, we consider hyphens and apostrophes as word "breaks". A tool that does not consider them as breaks would treat "l'émetteur-récepteur" as one word rather than three and get a different result.

2.5.2.1 Candidate A

Table 5: Operations needed to transform candidate A into reference.

Operation	Edited words	Number of editing steps
Matches	batterie, est, l', émetteur, récepteur	5
Shifts		0
Substitutions	le/la, fixé/installée, à/sur	3
Insertions		0
Deletions	bloc	1

Let us now compute the TER for candidate translation A:

$$\mathrm{TER}_A = \frac{0+3+0+1}{8} = 0.5 = 50\% \tag{7}$$

[13] Note that our examples contain no shifts, so that TER equals WER here.

2.5.2.2 Candidate B

Table 6: operations needed to transform candidate B into reference.

Operation	Edited words	Number of editing steps
Matches	la, batterie, est, l', émetteur, récepteur	6
Shifts		0
Substitutions	fixée/installée, à/sur	2
Insertions		0
Deletions		0

Let us now compute the TER for candidate translation B:

$$\text{TER}_B = \frac{0+2+0+0}{8} = 0.75 = 75\% \tag{8}$$

2.5.2.3 Candidate D

Table 7: operations needed to transform candidate D into reference.

Operation	Edited words	Number of editing steps
Matches	est, l', émetteur, récepteur	3
Shifts		0
Substitutions	le/la, bloc/batterie, fixé/installée, à/sur	4
Insertions		0
Deletions	piles	1

Let us now compute the TER for candidate translation D:

$$\text{TER}_D = \frac{0+4+0+1}{8} = 0.63 = 63\% \tag{9}$$

Systems A's output gets a TER score of 50%, system B's a TER score of 75% and system D's a TER score of 63%. Because WER and TER are error rates, these metrics take mismatches, and not matches, into account. This means that contrarily to the precision, recall and F-measure scores, the lower the value, the better the MT output is deemed to be. Thus, according to this metric, the best output would be candidate A.

2.5.3 Human translation edit rate (HTER)

Because the candidate translation could be acceptable while still being quite different from the reference translation, a metric that assesses the former against the latter could be unfairly harsh on the MT system. Alternatively, we can compute the *human translation edit rate* (HTER): in this variant, we ask a human evaluator to post-edit a candidate MT output, and we count the number of editing steps needed to transform that candidate output (and possibly other candidates) into the post-edited version (Snover et al. 2006).

Let us go back to our example. Compare candidate translations A, B and D with a post-edited segment, which could be any of the candidate translations, edited by a human reviser (Figure 8): what is the minimum number of steps required to go from candidates A, B and D to the post-edited segment?

Source text: Battery pack is attached to the transceiver.

Candidate A (eTranslation): Le bloc-batterie est fixé à l'émetteur-récepteur.

Candidate B (Google Translate): La batterie est fixée à l'émetteur-récepteur.

Candidate D (DeepL): Le bloc-piles est fixé à l'émetteur-récepteur.

Post-edited segment: La batterie est fixée à l'émetteur-récepteur.

Figure 8: Source text, candidate translations A, B and D, and post-edited segment

Again, remember the number of matches will not go into the final calculus.

2.5.3.1 Candidate A

Table 8: operations needed to transform candidate A into post-edited segment.

Operation	Edited words	Number of editing steps
Matches	batterie, est, à, l', émetteur, récepteur	6
Shifts		0
Substitutions	le/la, fixé/fixée	2
Insertions		0
Deletions	bloc	1

Let us now compute the HTER for candidate translation A:

$$\text{HTER}_A = \frac{0+2+0+1}{8} = 0.38 = 38\% \tag{10}$$

2.5.3.2 Candidate B

Table 9: operations needed to transform candidate B into post-edited segment.

Operation	Edited words	Number of editing steps
Matches	la, batterie, est, fixée, à, l', émetteur, récepteur	8
Shifts		0
Substitutions		0
Insertions		0
Deletions		0

Let us now compute the HTER for candidate translation B:

$$\text{HTER}_B = \frac{0+0+0+0}{8} = 0 = 0\% \tag{11}$$

2.5.3.3 Candidate D

Table 10: operations needed to transform candidate D into post-edited segment.

Operation	Edited words	Number of editing steps
Matches	est, à, l', émetteur, récepteur	4
Shifts		0
Substitutions	le/la, bloc/batterie, fixé/fixée	3
Insertions		0
Deletions	piles	1

Let us now compute the HTER for candidate translation D:

$$\text{HTER}_D = \frac{0+3+0+1}{8} = 0.5 = 50\% \tag{12}$$

Systems A's output gets an HTER score of 38%, system B's an HTER score of 0% and system D's an HTER score of 50%. Remember, because TER and HTER are error rates: the lower the value, the better the MT output is deemed to be. Thus, according to this metric, the best output would be candidate B.

Table 11: TER and HTER scores for each candidate translation

Metric	Candidate A	Candidate B	Candidate D
TER	50%	25%	63%
HTER	38%	0%	50%

Now, compare the TER and HTER scores for each candidate translation in our example (Table 11): they all get a lower, i.e. better, HTER than TER. Our example confirms that "the edit rate between a machine translation and its postedited version is dramatically lower than between the machine translation and an independently produced human reference translation" (Koehn 2020: 52). This difference could be taken as a reminder of the dangers of under-post-editing, which can happen when post-editors work under too much time pressure.[14]

2.5.4 Bilingual evaluation understudy (BLEU)

The *bilingual evaluation understudy* (BLEU) metric represents the *n*-grams shared between a candidate machine translation and a reference translation.[15] Therefore, it takes both the number of matching words and word order into account. It is usually set to consider *n*-grams from unigrams to 4-grams and can allow for different *n*-grams to be weighted differently.

Figure 9 presents a candidate and a reference translation for the sentence made famous by René Magritte's painting *The Treachery of Images*. The candidate translation shares five out of a possible six 1-grams with the reference, if we count the punctuation mark at the end of the sentence as a 1-gram. The candidate also contains four out of the five 2-grams in the reference ([is not], [not a], [a pipe], [pipe .]), and three out of the four 3-grams in the reference ([is not a], [not a pipe], [a pipe .]). Finally, it contains two of the three 4-grams in the reference. Here, the overlapping 4-grams are [is not a pipe] and [not a pipe.], the first of which is illustrated in Figure 9.

[14]The authors would like to thank Mikel Forcada for this comment.

[15]As mentioned above, note that BLEU can also allow for the use of multiple reference translations.

Source text:	Ceci n'est pas une pipe.
Candidate Translation (fictional):	That is not a pipe .
Reference:	This is not a pipe .

Figure 9: candidate and reference translation for the sentence Ceci n'est pas une pipe, showing one 4-gram overlap.

Because the BLEU score computes the ratio of *n*-grams in the candidate translation that also occur in the reference translation, it is a precision metric. Table 12 thus presents the precision scores (expressed as a ratio and a decimal fraction) for each order of *n*-gram in our candidate translation.[16]

Table 12: Precision (from 1-grams to 4-grams) for the candidate translation 'That is not a pipe'.

Metric		
Precision (1-gram)	5/6	0.83
Precision (2-gram)	4/5	0.80
Precision (3-gram)	3/4	0.75
Precision (4-gram)	2/3	0.66

To compute an overall BLEU score for a candidate translation, we compute the geometric mean (a special type of "average") of these individual precision scores. It turns out to be just under 0.76 or 76%.[17] (This is actually a very high BLEU score, but then the example was a very simple one.)

It should be noted, however, that BLEU scores are usually computed over entire corpora, not individual sentences. And because precision metrics can be tricked by systems that produce translations only for words the system is sure of (think of the student who refuses to name more days of the week, in case they get any wrong), BLEU uses a *brevity penalty*. The brevity penalty is the ratio between the number of words in the candidate translation and in the reference translation (for more on this, see Koehn 2020: 227). It kicks in when the candidate translation is shorter than the reference translation. No brevity penalty is imposed in the case

[16]The presentation used in Table 12 is borrowed from (Koehn 2010: 227).

[17]Readers can use familiar spreadsheet software to calculate geometric means. A specifically-designed *BLEU score calculator* may also need to take into account any extra weight for, e.g., longer *n*-grams, and should have a way of *smoothing* out any zeros that appear in the *n*-gram precision scores. For details, see Post (2018).

of the sentence in Figure 9, however, as the candidate is the exact same length as the reference.

Although this metric is often referred to as "the BLEU score", there are so many different parameters that go into computing BLEU scores (Post 2018) that it can be very difficult for non-specialists to find out and understand how exactly a given AEM tool computes it. What is important for translators who wish to use this AEM to assess MT outputs is that the scores they get for different candidate translations are consistently calculated: put simply, make sure you use the same MT evaluation tool, that you understand the settings it uses, and, if some of them are user-definable, that you use the same settings when comparing candidate translations using that AEM. This way, you will get comparable scores.

By way of illustration, Table 13 shows the BLEU scores computed by two different calculators, those provided by MutNMT and Tilde, for the candidate translations in Figure 4.[18]

Table 13: Sentence-level BLEU scores for candidate translations A, B and D using MutNMT and Tilde

BLEU Calculator	Candidate A	Candidate B	Candidate D
MutNMT	15%	31%	15%
Tilde	50%	61%	47%

According to these scores, system B's output is better than system A's. This is consistent with our findings so far. But the actual values vary dramatically and the user would need to investigate why this might be the case.

2.5.5 ChrF3

The ChrF score is an F-measure based on character *n*-grams. Therefore, it is based both on precision and recall. Remember the formula for the F-measure:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{13}$$

The formula for the ChrF score is:

$$\text{ChrF}\beta = \left(1 + \beta^2\right) \cdot \frac{\text{ChrP} \cdot \text{ChrR}}{\beta^2 \cdot \text{ChrP} + \text{ChrR}} \tag{14}$$

[18] https://mutnmt.prompsit.com/index; MutNMT uses the SacreBLEU algorithm (Post 2018). Tilde's "interactive bleu score evaluator" is available at https://www.letsmt.eu/Bleu.aspx.

where

- ChrP is the character *n*-gram precision, i.e. the number of correct character *n*-grams in the candidate translation divided by the total number of *n*-grams in the candidate translation,
- ChrR is the character *n*-gram recall, i.e. the number of correct character *n*-grams in the candidate translation divided by the total number of character *n*-grams in the reference translation, and
- β is a parameter which assigns β times more importance to recall than to precision. If $\beta = 1$, then they are equally important; ChrF1 is the harmonic mean of character *n*-gram precision and recall (Popović 2015).[19]

The ChrF3 score, then, is a variant of ChrF where $\beta = 3$, i.e. recall has three times more weight than precision. According to Popović (2015), experiments have shown that ChrF, and especially ChrF3, represent promising metrics for automatic evaluation of MT output.

As with BLEU, we will not calculate ChrF3 scores here. However, it is interesting to compare the scores given by an AEM tool.

Table 14 shows the ChrF3 scores of candidates A, B and D, as computed by MutNMT.

Table 14: ChrF3 score for candidate translations A, B and D

Metric	Candidate A	Candidate B	Candidate D
ChrF3	64%	69%	49%

According to these scores, system B's output is once again rated better than system A's and system C's.

2.5.6 A word of caution: AEM scales

To end this section on AEMs, we would like to sound a note of caution: when dealing with AEM scores, make sure you know what the numbers mean. Some scores can be reported as decimals or as percentages (e.g. 0.8 or 80%). And on a scale from 0 to 1 (or 0% to 100%), 0 could be the best score and 1 the worst for one

[19]Note that this also applies to F: thus far, all F-measures were F_1 measures where $\beta = 1$. The value of β may be changed.

metric (e.g. TER) while 1 could be the best score and 0 the worst for another (e.g. BLEU).

We have also seen that caution should be exercised when comparing metrics computed with different tools: indeed, the algorithms behind apparently similar AEMs might differ in ways the non-specialist user is unaware of.

Finally, it should be said that to make the most of these different metrics, users will need to reflect on what the scores mean for their purposes and get used to them. Comparing different AEMs and combining them with human evaluation will help, even though you might be faced with differences (Doherty 2017: 134). Human evaluation can take into account context in a better way, provided that it is not done with segments presented in a random order: evaluators can then look for errors in pronouns, for instance, while AEMs mostly operate at word and sentence level. One interesting way to combine measures is to use both HTER, which gives you a measure of technical post-editing effort, and a temporal measure, which tells you how long post-editing took.[20] O'Brien (2022 [this volume]) gives an overview of measures of post-editing effort.

2.6 Type-token ratios

The *type-token ratio* (TTR) is not quality metric as such. Rather, it provides a global insight on the lexical variety of a text. It is mentioned here because it has become one of the metrics used to comment on ways in which machine translated text can differ from other texts in the same language (see Toral 2019), and some evaluation interfaces, including that of MutNMT, are now reporting this metric.

TTR basically measures vocabulary variation (or *lexical variety*, Williamson 2009) within a text or corpus. The number of running words in a text is referred to as the number of *tokens*. But words can be repeated in a text; if the same word occurs three times in a text, for example, it counts as three tokens, but only as one *type*. The relationship between number of tokens and number of types is called type-token ratio and is computed as follows:

$$\text{type token ratio} = \frac{\text{no. of types}}{\text{no. of tokens}} \tag{15}$$

or

$$\text{type token ratio} = \frac{\text{no. of types}}{\text{no. of tokens}} \cdot 100 \tag{16}$$

[20] While averaging them into a single score might not be very telling, looking for correlations could help to identify the most serious problems.

The first method gives results ranging from 0 to 1, while the second gives percentages, ranging from 0% to 100%. The higher the type-token ratio, the more varied the vocabulary in the text under scrutiny.

However, several warnings have to be issued regarding this metric. Firstly, TTR is highly sensitive to text-length. Indeed, the longer a text is, the more often such words as determiners and articles will be repeated. Moreover, because texts, especially specialized texts, have a thematic unity, terms are repeated. Therefore, the longer the text segment under consideration, the lower the TTR. Because of this sensitivity of TTR to text length, TTR may have to be standardized across blocks of a given number of tokens (e.g. 1,000 tokens) depending on the task at hand. Standardizing in this way would allow you to compare the TTR of your machine translated corpus with that of a corpus of different length in the same (target) language.

Secondly, while lemmatization does not matter when comparing texts in the same language, TTR has to be lemmatized when comparing two or more languages as some languages have richer inflectional morphology than others and thus would be expected to have more lexical variety, simply because they have, for example, more forms for any given verb. If you are simply using standardized TTRs to compare the lexical variety of machine translated texts with that of other texts in the same language however, then lemmatization will not be necessary.

Lastly, bear in mind that a higher TTR, that is, one that indicates more lexical variety, does not necessarily equate with higher complexity. For example, consider the sentences "The girl saw a fire." and "The lexicographer observed the conflagration." Both sentences are made up of five words (tokens), but while the former has five types, the latter has only four (because the token "the" occurs twice). The first sentence thus has a TTR of 1 or 100%, while the second has a TTR of 0.8 or 80%. But in spite of being less varied than the first sentence, the second sentence is more complex.[21]

As already indicated, segment-level comparisons of TTRs might not make much sense, but at text or corpus level, same-language comparisons of standardized TTRs could give us valuable information, depending on the kind of text we are dealing with. This chapter has focused on specialized translation. But there are different kinds of specialized translation, which follow different conventions. Contrary to literary or marketing translation, where higher lexical variety (and thus a higher TTR) could be associated with higher quality and make the reading all the more pleasant for the user of the target text, technical translation often has to comply with certain conventions that tend to decrease the lexical variety of

[21]The authors would like to thank Dorothy Kenny for this comment.

texts while making them easier to use for the end user. The main example used in this chapter comes from a user manual, which means in turn that it should follow such conventions as using a single term for a single concept, with no variation, and that instructions should as far as possible be written following the same pattern. For example, if "transceiver" was translated at times by "émetteur-récepteur", and at others by "radio" and by "appareil", it would lead to a higher TTR, while introducing uncertainty for the end user.

This leads us to conclude this section with a second word of caution.

3 Conclusion

In this chapter we have sought to illustrate what a pragmatic approach to MT evaluation implies for specialized translators or trainees. This approach has been called pragmatic because it considers evaluation as a means to an end, and implies choosing among different methods depending on the situation, often using a combination of human and automatic evaluation.

While the comparison of MT outputs has been used as a method throughout this chapter, it is worth noting that specialized translators are rarely given a choice about what evaluation metric to use in current translation scenarios. Rather, they often need to make a quick judgement on whether a given MT solution is fit for purpose, or provide a general assessment of its quality.

We have thus explained how evaluations of MT outputs might be conducted, using a combination of human and automatic evaluation metrics. We have explained the latter in great detail because we believe that, for all their limitations, they can be put to good use if understood properly, and combined with human evaluation.

References

Aziz, Wilker, Sheila Castilho & Lucia Specia. 2012. PET: a tool for post-editing and assessing machine translation. In *Proceedings of the eight international conference on language resources and evaluation (LREC'12)*, 3982–3987.

Castilho, Sheila. 2020. On the same page? Comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In *Proceedings of the 5th conference on machine translation (WMT)*, 1150–1159. https://aclanthology.org/2020.wmt-1.137.pdf.

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.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley & Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics* 108. 109–120. DOI: 10.1515/pralin-2017-0013.

Castilho, Sheila & Sharon O'Brien. 2016. Evaluating the impact of light post-editing on usability. In *10th international conference on language resources and evaluation (LREC)*, 310–316. May 2016, Portorož, Slovenia. ELRA.

Doherty, Stephen. 2017. Issues in human and automatic translation quality assessment. In Dorothy Kenny (ed.), *Human issues in translation technology*, 131–148. London: Routledge.

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

Gouadec, Daniel. 2010. Quality in translation. In *Handbook of translation studies. Volume 1*, 270–275. John Benjamins Publishing Company.

Grbić, Nadja. 2008. Constructing interpreting quality. *Interpreting* 10(2). 232–257.

House, Juliane. 2015. *Translation quality assessment: Past and present*. London: Routledge.

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.

Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Koehn, Philipp. 2020. *Neural Machine Translation*. Cambridge: Cambridge University Press.

Mariana, Valerie, Troy Cox & Alan Melby. 2015. The multidimensional quality metric (MQM) framework: A new framework for translation quality assessment. *The Journal of Specialised Translation* 23. 137–161.

Moorkens, Joss. 2018. What to expect from neural machine translation: a practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer* 12(4). 375–387.

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.

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.

Popović, Maja. 2015. Chrf: Character n-gram f-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, 392–395. Association for Computational Linguistics. 10.18653/v1/W15-3049.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the third conference on machine translation (WMT), Volume 1: research papers*, 186–191. Association for Computational Linguistics. DOI: 10.18653/v1/W18-6319.

Qin, Ying & Lucia Specia. 2015. Truly exploring multiple references for machine translation evaluation. In *Proceedings of the 18th annual conference of the European Association for Machine Translation*, 113–120. https://aclanthology.org/W15-4915/.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla & John Makhoul. 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*, 223–231. Cambridge, Massachusetts: Association for Machine Translation in the Americas. https://aclanthology.org/2006.amta-papers.25/.

Toral, Antonio. 2019. Post-editese: An exacerbated translationese. In *Proceedings of machine translation summit XVII*, 273–281. EAMT. https://www.aclweb.org/anthology/W19-6627/.

Williamson, Graham. 2009. *Type-token ratio*. Last retrieved 5 Dec. 2020. https://www.sltinfo.com/wp-content/uploads/2014/01/type-token-ratio.pdf.

Yamada, Masaru. 2019. The impact of Google neural machine translation on post-editing by student translators. *The Journal of Specialised Translation* 31(2019). 87–106.