From terminological inconsistencies to major mistranslations. A qualitative analysis of NMT errors in public health communication

Vanessa Šorak (Heidelberg University)

vanessa.sorak(at)iued.uni-heidelberg.de

Abstract

This article presents a qualitative analysis of frequent NMT errors in public health communication. Its primary aim is to sensitise users and decision-makers to common issues and raise public awareness for potential pitfalls and limits of current state-of-the-art NMT systems. In this context, the article also addresses the usability of raw NMT output in emergency situations. The investigation itself focuses on pandemic-related WHO texts and consists of a fine-grained manual error analysis encompassing three languages (English, French, and Spanish) and two NMT systems (DeepL and Google Translate). The five most frequent error types observed in this investigation included mistranslations, inconsistent use of terminology, unidiomatic or awkward style, untranslated text, and other internal inconsistencies. These error types are illustrated with examples and analysed in terms of their severity, their underlying causes, and their potential consequences. The findings show that raw NMT output is useful only to a very limited extent and that the risks and benefits associated with its use should be assessed extremely carefully.

Keywords

Neural Machine Translation, public health communication, COVID-19, Machine Translation Literacy, error analysis

1 Introduction

The COVID-19 pandemic drew urgent attention to the need for fast and accurate translation in emergency situations, sparking a debate about the untapped potentials of Neural Machine Translation (NMT). One major point of discussion is the usability of raw (i.e. unedited) NMT output, which could significantly accelerate the dissemination of time-sensitive health information. However, this also entails some serious risks: past research has shown that NMT is highly susceptible to subtle, yet serious errors that are difficult to identify for untrained users (e.g., Isabelle et al. 2017). This can be particularly dangerous in the medical domain, where even minor mistranslations may have dramatic consequences. It is therefore critical for users and relevant decision-makers to acquire a certain degree of “MT literacy” (Bowker and Ciro 2019), i.e. the ability to adequately assess the quality of raw MT output as well as the risks involved in using it. Familiarising oneself with typical errors and challenges through error analysis can be particularly useful in this regard, as it offers valuable insights into the strengths and weaknesses of state-of-the-art NMT systems.

The present article aims to make a small contribution in this area by addressing the most important challenges and deficiencies of NMT in the context of public health communication. It presents the results of a detailed manual error analysis conducted with two state-of-the-art NMT systems (DeepL and Google Translate) and three languages (English as source language and French and Spanish as target languages). This analysis was carried out as part of a pilot study appertaining to an ongoing doctoral research project on the risks and potentials of NMT in pandemic-related health communication. The pilot study consisted of a comprehensive corpus analysis based on a triangulation approach including both quantitative and qualitative methods. Conclusions drawn from the quantitative part are presented in Šorak (2025) and will not be further discussed in this article. Instead, it will focus on the qualitative aspects of the investigation, delving further into detail on the most frequent error types and systematically analysing their underlying causes. The primary objective is to sensitise users and decision-makers to these issues and raise awareness for the challenges and limits of current state-of-the-art NMT systems.

2 Related Work

There are numerous studies related to the quality of NMT in health communication, yet only a few go into detail regarding frequent error types and their causes. Most papers employ error analysis to evaluate the general translation quality (cf. Zappatore and Ruggieri 2024: 26, 31–34) and while they typically provide some illustrative examples, they seldomly discuss the identified errors systematically. Studies that extend the analysis to the underlying error sources appear to be distinctly rare, even though this information may be particularly relevant for users trying to assess if a certain text is suitable for NMT. Related studies that incorporate the underlying error sources at least to some degree are Khoong et al. (2019), Almahasees et al. (2021), and Pym et al. (2022) who investigated the use of Google Translate in clinical settings as well as public health communication. Khoong et al. focused on translations of ER instructions for English to Spanish and Chinese, while the other two papers analysed errors in COVID-19 related texts for English to Arabic as well as Catalan to English and French. All studies found that while the translation quality was generally high, the output still contained some major errors with a considerable potential for health risks. To the best of our knowledge, there are to date no other papers in this domain dedicated to systematically analysing frequent NMT error types vis-à-vis their causes.[1] There furthermore seems to be a pronounced lack of studies that conduct error analyses on document-level. However, given the fact that current NMT systems are typically unable to incorporate more than a few sentences as context and therefore take translation decisions locally, they often struggle to produce consistent and coherent translations. As shown among others by Läubli et al. (2018), taking the text as a

[1] Comprehensive overviews of research related to NMT in health communication can be found in Vieira et al. (2020), Haddow et al. (2021), and Zappatore and Ruggieri (2024).

whole into account to uncover issues related to internal inconsistencies, text coherence, and translation strategy can thus have a significant impact on the assessment and perception of NMT quality. The present article tries to address these research desiderata by presenting findings from a detailed, document-level analysis including not only the error type and severity but also the underlying error causes.

3 Data and Method

The investigation focused on translations of COVID-19 related WHO texts from English into French and Spanish generated by Google Translate and DeepL Pro. As already stated above, it was carried out within the scope of a mixed methods pilot study from an ongoing doctoral research project. The present article provides insights into the qualitative aspects of this pilot study while conclusions related to the quantitative part are presented in Šorak (2025). The main focus of the present work lies on the five most frequent error categories, which will be discussed vis-à-vis their underlying causes and illustrated with examples. The article will also examine some of the potential risks and consequences associated with these errors.

3.1 Data Selection

The texts selected for this pilot study appertain to a larger corpus comprising 200 English source texts and their official translations into French and Spanish. All texts were extracted from the WHO online archive *Institutional Repository For Information Sharing* (IRIS) within the framework of the above-mentioned doctoral research project. The corpus exclusively contains texts related to the COVID-19 pandemic published between January 2020 and August 2022. For the smaller scale pilot study, a subcorpus comprising 5 different text genres was created. These text genres were selected based on (a) their relevance, i.e. their relative frequency in the main corpus, and (b) their main characteristics, including the degree of specialisation,[2] the syntactic complexity,[3] and other attributes such as extensive formatting, continuous vs. elliptical text, bulleted lists, etc.[4] The main purpose of incorporating these aspects was to achieve a diverse subcorpus that includes as many different translational challenges as possible. All documents were translated in their entirety (as .docx files) into French and Spanish using DeepL Pro and Google Translate. One document per text genre was then chosen for the manual error analysis, leading up to a total of 20 machine-generated translations comprising 90,197 tokens. The official WHO translations served as references for the analysis.

[2] Determined by the density of technical terms as well as the target audience (experts vs. laypeople).

[3] Determined on the basis of the Flesch-Kincaid readability test (Microsoft 2023).

[4] More details on these characteristics as well as their impact on overall translation quality can be found in Šorak (2025). The article also provides an analysis of the text genres' suitability for MT.

3.2 Error Analysis

The error metric presented below was specifically designed for the purpose of analysing the risks and potentials associated with the use of state-of-the-art NMT systems in pandemic-related health communication. It is a modified version of the DQF-MQM metric (Lommel et al. 2014)[5] and encompasses 7 categories and 21 subcategories:

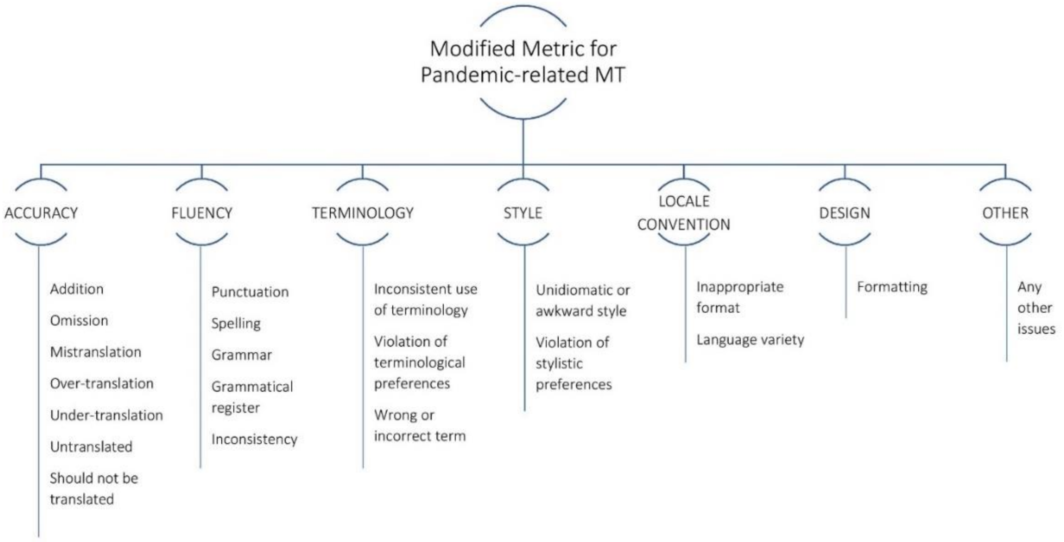

Figure 1: The modified error metric for pandemic-related NMT (based on DQF-MQM).

Using DQF-MQM as a basis for this metric presented several advantages: firstly, it is one of the most widely-used error metrics in MT research and has been continuously updated and improved to reflect best practices in research as well as current industry standards (cf. Lommel et al. 2015: 4f.).[6] Comprising 8 categories and 34 subcategories itself, it is also a comparatively detailed metric, allowing for a fine-grained error analysis at both the segment and document level. DQF-MQM furthermore includes several subcategories aimed at capturing subtle errors such as semantic nuances (e.g. over- and undertranslations) and issues related to textual or terminological consistency, which makes it particularly suitable for high-quality NMT. Last but not least, it provides detailed definitions and examples for all categories, subcategories, and severity levels, which facilitates the annotation process and improves consistency. However, the metric was still missing some important subcategories while also encompassing several other categories that were

[5] Available at https://info.taus.net/dqf-mqf-error-typology-template-download [19.08.2023].

[6] DQF-MQM was updated in 2021 and has since been available under the name MQM Core on the World Wide Web Consortium website: https://www.w3.org/community/mqmcg/updated-mqm-core-posted/ [19.08.2023]. The main error categories remained essentially the same, but some subcategories were modified and others added. Given the fact that a higher number of categories typically leads to lower annotation consistency and that MQM Core is even more detailed than DQF-MQM, the latter was chosen over the updated version.

irrelevant for this investigation. Following best practices as defined inter alia by Popović (2018),[7] it was therefore decided to design a modified version that is specifically tailored to the requisites of the present study. The modifications included three additions, eleven eliminations, and several mergers of subcategories. The severity levels (cf. Table 1)[8] as well as the category definitions were largely adopted from the original metric, although some definitions needed to be slightly altered to reflect the above-mentioned modifications.

Table 1: Severity levels used in this analysis as defined by DQF-MQM.

Severity	**Definition**
Major	Errors that may confuse or mislead the user or hinder proper use of the product/service due to significant change in meaning or because errors appear in a visible or important part of the content.
Minor	Errors that don't lead to loss of meaning and wouldn't confuse or mislead the user but would be noticed, would decrease stylistic quality, fluency or clarity, or would make the content less appealing.
Neutral	Used to log additional information, problems or changes to be made that don't count as errors, e.g. they reflect a reviewer's choice or preferred style, […] or instruction/glossary changes not yet implemented […].

Discussing every single category and subcategory in detail would exceed the limits of this article, which is why the following sections will focus on the five most frequent error subcategories as well as the factors that caused them.[9] If a segment contained more than one error or a combination of errors (e.g. INCONSISTENT USE OF TERMINOLOGY and MISTRANSLATION), they were annotated separately. Repeated errors on the other hand were provisionally excluded from the analysis.

It should also be noted in this context that the term "error" is used rather broadly here, as it encompasses issues that are not necessarily *incorrect*: the category VIOLATION OF TERMINOLOGICAL PREFERENCES for instance captures deviations from standard WHO terminology that are otherwise perfectly acceptable. The fact that these issues are not necessarily errors is accounted for by the severity level NEUTRAL (cf. Table 1). Since the primary purpose of this article is to sensitise users for the most important risks associated with raw NMT output, it will exclude these and all other NEUTRAL issues.

The error metric was furthermore extended to incorporate an analysis of the underlying error cause. All error causes were identified during the annotation process itself, i.e. they were not predetermined but collected based on observation.

[7] Note that it is generally recommended to adapt error metrics to the individual task at hand, as research purpose, language pairs, domain, and text genres heavily influence the type, frequency and severity of errors that may be occurring (cf. Popović 2018: 12).

[8] DQF-MQM includes an additional severity level (Critical) which was omitted from the Table because no such errors were identified in this corpus. It also offers a Kudos feature for commendable translations which will not be discussed here as the focus of this paper lies solely on errors.

[9] An overview of all error categories and their definitions can be found in Šorak (2025).

If the error cause couldn't be identified beyond a reasonable doubt, it was annotated as UNCLEAR.[10]

4 Results

The investigation revealed that the most important error types in terms of frequency were MISTRANSLATION and VIOLATION OF TERMINOLOGICAL PREFERENCES, accounting for 19% and 16% of all annotated errors (761 in total). As already stated above, the latter will be disregarded in the ensuing analysis since they are not directly relevant for this paper: typically annotated as neutral, these issues have no significant impact on comprehensibility, provided that the correct alternative to the WHO term is used consistently throughout the translation. All other subcategories were considerably less prevalent, with INCONSISTENT USE OF TERMINOLOGY being the third most frequent at 9% and UNIDIOMATIC OR AWKWARD STYLE and UNTRANSLATED TEXT tied for fourth place at 8% each. The category INCONSISTENCY, which captures inconsistencies of non-terminological nature, was the fifth most frequent at 7%. The relative frequency of all error subcategories identified within the corpus is illustrated by the figure below:

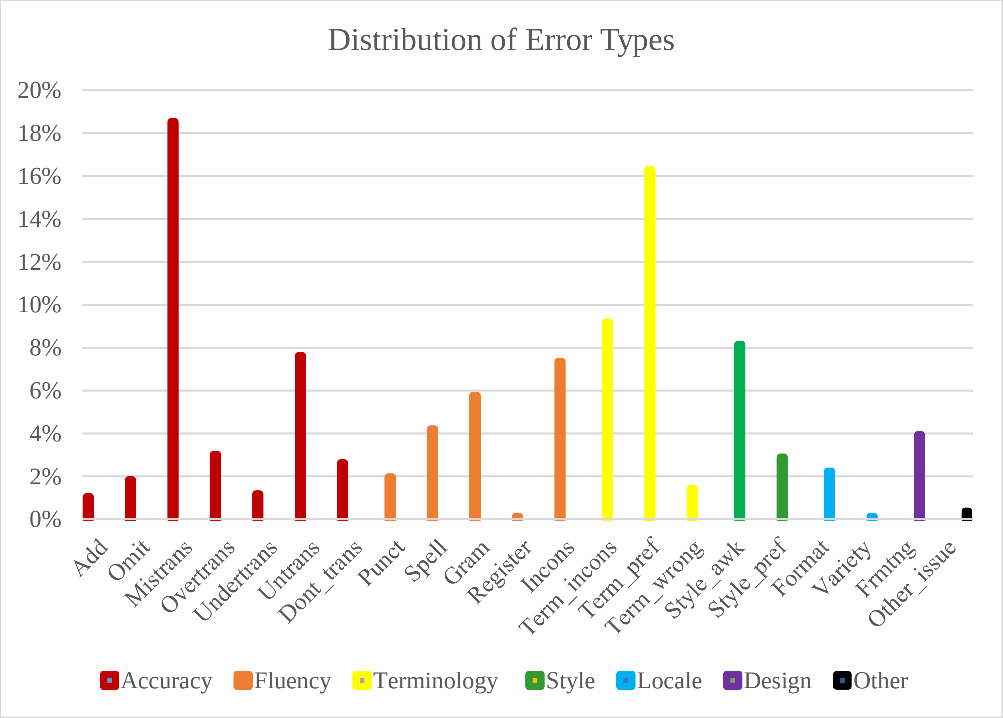

Figure 1: Relative frequency of all error subcategories identified.

[10] It should be noted that all error causes were identified based on plausibility only and are therefore to be interpreted as *presumed* error causes. In cases where several risk factors could have caused the error, the more prominent factor was chosen. An overview of all identified error causes can be found in Šorak (2025).

The five most prevalent categories will be analysed in more detail and illustrated with examples in the following subsections. This is followed by an additional example showing a combination of several error types.

4.1 Mistranslations

The category MISTRANSLATION captures instances where the semantic content of the source text is misrepresented or not accurately reflected in the target text. Mistranslations appear to be predominantly caused by ambiguity, with syntactic or referential ambiguity accounting for 29% and lexical ambiguity for 26% of all annotated cases (141 in total). Technical terms were identified as the third most frequent cause at 16%. The examples below illustrate these three error causes:

(1) Mistranslation caused by syntactic ambiguity (110_30_D_FR)

ST: Chest radiography: hazy opacities, often rounded in morphology, with peripheral and lower lung distribution

REF: Radiographie thoracique : opacités à contours flous, souvent de morphologie arrondie, réparties à la périphérie et dans la partie inférieure des poumons

DeepL: Radiographie du thorax : opacités brumeuses, souvent de morphologie arrondie, avec une distribution périphérique et inférieure des poumons.

(2) Mistranslation caused by lexical ambiguity (145_286f._G_FR)

ST: Influenza virus: □ Positive □ Negative □ Not done
If positive, type:

REF: Virus grippal : □ Positif □ Négatif □ Non réalisé
Si positif, type:

Google: Virus de la grippe : □ Positif □ Négatif □ Pas fait
Si positif, tapez:

(3) Mistranslation caused by a technical term (145_94_D_FR)

ST: Lower chest wall indrawing □ Yes □ No □ Unk

REF: Tirage sous-costal □ Oui □ Non □ Inc

DeepL: Dessin de la paroi thoracique inférieure □ Yes □ Non □ Unk

Ex. 1 shows a mistranslation involving a complex nominal phrase, namely “hazy opacities […] with peripheral and lower lung distribution” which was translated into French as “with a peripheral and lower distribution of the lungs”. Complex nominal phrases in English are often challenging for NMT because the semantic relations and syntactic dependencies of their constituents are not explicitly expressed. While this typically doesn't pose a problem for human readers, NMT systems often fail to correctly disambiguate these phrases. The same is true for lexical ambiguities such as in ex. 2, where the noun "type" was misinterpreted and translated as the homonym verb in the imperative mood. The third example shows a mistranslation caused by a technical term, namely "lower chest wall indrawing" which was translated as "drawing [in the sense of a picture] of the lower thoracic wall". It is likely that the original term wasn't represented frequently enough in the NMT system's training data and was therefore translated with the help of byte-pair encoding.[11] If this is the case, the mistranslation could be explained by the orthographic similarity of *indrawing* and *drawing*.

Given the fact that mistranslations per definitionem involve a change or loss of meaning, they are by and large severe errors: accordingly, 73% of all identified mistranslations in this analysis were classified as MAJOR errors. Major mistranslations can lead to an impaired comprehensibility as well as the dissemination of false information. In particularly grave cases, this may also result in serious repercussions, ranging from reputational or financial damages as far as human injury and even death.[12]

4.2 Inconsistent use of terminology

The category INCONSISTENT USE OF TERMINOLOGY is defined as the use of multiple terms for the same concept within one text in cases where consistency is desirable. This encompasses both the use of correct terminological alternatives and incorrect terminology. The category thus frequently occurs in combination with other issues such as VIOLATION OF TERMINOLOGICAL PREFERENCES and MISTRANSLATION. This is also reflected by the underlying error causes: almost half of all identified cases (70 in total) were caused by the use of a correct terminological variant (44%) and almost a fourth by inconsistent translations of proper names (24%). Technical terms only accounted for approximately 6%, which is primarily due to the fact that only a single error source was annotated per error: if the terminological inconsistency involved a technical term, but resulted from the use of a correct alternative, the error source was annotated as TERMINOLOGICAL VARIANT (cf. ex. 4). The 6% mentioned above hence represent instances where the inconsistency was exclusively related to an *incorrect* translation of a technical term, e.g. a mistranslation as seen in ex. 3. The same applied to (medical) acronyms. The mistranslation of a medical acronym in combination with the inconsistent use of terminology is illustrated in Section 4.6 (ex. 13). The following examples show the two primary causes of terminological inconsistency:

[11] Lexical items that are rare or unseen in the training data are commonly referred to as Out-Of-Vocabulary (OOV) words. Typical examples of OOVs include proper names (cf. ex. 5) and highly technical terms. Byte-pair encoding (i.e. splitting the unknown word into sub-word units) is one of the most-widely used techniques to address this issue (cf. Araabi et al. 2022: 117).

[12] According to the DQF-MQM metric, such errors would be classified as *critical*. However, no errors of this nature were identified in this analysis.

(4) Terminological inconsistency caused by a correct terminological variant (145_89_140_D_ES)

ST: Angiotensin II receptor blockers (ARBs)? ☐ Yes ☐ No ☐ Unknown

REF: Antagonistas de los receptores de la angiotensina II (ARA-II) ☐ Sí ☐ No ☐ Se desconoce

DeepL: Bloqueantes de los receptores de la angiotensina II (ARA) ☐ ☐ Sí ☐ No ☐ Sin datos (seg. 89)

Antagonistas de los receptores de la angiotensina II (ARA) ☐ Sí ☐ No ☐ Sin datos (seg. 140)

(5) Terminological inconsistency caused by a proper name (080_1_4_D_FR)

ST: Country COVID-19 intra-action review (IAR)

REF: Revue intra-action (RIA) de la COVID-19

DeepL: Examen intra-action (IAR) de la COVID-19 par pays (seg. 1)

[...] revue intra-action (RIA) COVID-19 du pays (seg. 4)

Ex. 4 shows an instance where terminological inconsistency was caused by the use of a correct terminological variant in the target language. This additionally led to a minor incoherence in segment 89, where the corresponding acronym to the full term would be BRA instead of ARA. Ex. 5 on the other hand shows an instance where the inconsistency was caused by a proper name: the “Country COVID-19 intra-action review” (IAR) is a WHO process for reviewing national response strategies to the pandemic. In this case, both the full term and the acronym were translated inconsistently.

Given the fact that most terminological inconsistencies were caused by the use of correct alternatives, it is hardly surprising that they were predominantly classified as MINOR issues (76%). It should be emphasised that this subcategory was only annotated in cases where consistency is desirable or necessary and that a certain lexical variety can of course be justified. However, terminological consistency can be extremely important in certain text passages (such as tables including repetitions of keywords) and in highly specialised or standardised texts. One example of such a text is the WHO “Case Report Form”. These documents are questionnaires designed for collecting clinical data on COVID-19 cases, including treatment, disease progression and outcome. The texts are characterised by a very high degree of specialisation due to a great density of medical terms, including acronyms and abbreviations. Most of these highly technical terms are repeated throughout the questionnaire to monitor disease progression. Given the fact that the primary aim is for clinicians to collect this data in an efficient and systematic way and thereafter report it to WHO for evaluation, it is paramount that technical terms are used and translated consistently throughout the entire document.

Terminological inconsistencies, even if due to a correct alternative, can lead to impaired comprehensibility, inefficient data collection, and possibly even biased or false results. It should be noted, however, that these negative effects are by no means limited to this specific text genre. Terminological consistency with regards to technical terms, acronyms or abbreviations can be equally important in less specialised texts, especially if they address the general public: in these cases, the use of several alternatives within one text may confuse or mislead laypeople readers who might not be aware that the different terms relate to the same concept.

4.3 Unidiomatic or awkward style

The category UNIDIOMATIC OR AWKWARD STYLE captures stylistic issues in the target text, i.e. text passages that are grammatically correct, but stylistically inadequate or unnatural. Unfortunately, it remained largely unclear what exactly causes unidiomatic or awkward style in the target language: in 35% of all annotated cases (63 in total), no specific reason could be identified. The most common identifiable cause was an interference of the source language (i.e. the adoption of source language structures), accounting for 27%. Redundancies in the source text also played a minor role (6%), often leading to awkward repetitions in the target text. This is illustrated in ex. 7, while ex. 6 shows the adoption of source language structures:

(6) Unidiomatic or awkward style caused by interference of the source language (110_34_G_FR)

ST: Clinical and public health judgment should be used to determine the need for further investigation in patients who do not strictly meet the clinical or epidemiological criteria.

REF: En el caso de los pacientes que no cumplan estrictamente los criterios clínicos o epidemiológicos, la decisión de realizar o no más exploraciones deberá basarse en un razonamiento de salud pública.

Google: Se debe utilizar el juicio clínico y de salud pública para determinar la necesidad de investigación adicional en pacientes que no cumplan estrictamente los criterios clínicos o epidemiológicos.

(7) Unidiomatic or awkward style caused by redundancy in the source text (145_55_G_ES)

ST: Symptom onset (date of first/earliest symptom)
[_D_][_D_]/[_M_][_M_]/[_2_][_0_][_Y_][_Y_]

REF: Inicio de los síntomas (fecha del primer síntoma)
[_D _][_D_]/[_M _][_M _]/[_2_][_0_][_A _][_A_]

Google: Inicio de los síntomas (fecha del primer/primer síntoma)
[_D _][_D_]/[_M_][_M_]/[_2_][_0_][_Y_][_Y_]

Issues related to style are inherently MINOR errors, as they do not encompass a change or loss of meaning. At first glance, they may therefore not appear to be particularly relevant, especially if compared to other issues such as severe mistranslations. It should be kept in mind, however, that public health communication involves many texts with an important appellative function. These texts are meant to convince the public to participate in vaccination programmes, help prevent the spread of infectious diseases, or build trust in health authorities. Awkward or inadequate style can weaken the acceptability and persuasiveness of these texts and ultimately lead to a negative impact on people's health decisions.

4.4 Untranslated text

The category UNTRANSLATED TEXT relates to content that should have been translated but was left untranslated in the target text. The primary cause for untranslated text were acronyms, accounting for 41% of all annotated cases (58 in total). Abbreviations were the second most frequent cause (28%), followed by heavy formatting of the source text (10%). These three factors are illustrated by the examples below:

(8) Untranslated text caused by acronyms (145_171_D_ES)

ST: APTT/APTR

REF: TTPa/índice de TPa

DeepL: APTT/APTR

(9) Untranslated text caused by an abbreviation (145_92_G_FR)

ST: SIGNS AND SYMPTOMS ON ADMISSION (Unk = Unknown)

REF: SIGNES ET SYMPTÔMES À L'ADMISSION (Inc = Inconnu)

Google: SIGNES ET SYMPTÔMES À L'ADMISSION (Unk = Inconnu)

(10) Untranslated text caused by heavy formatting (145_363_D_FR)

ST: Outcome: ☐ Discharged alive ☐ Hospitalized ☐ Transfer to other facility ☐ Death ☐ Palliative discharge ☐ Unknown

REF: Issue : ☐ Sortie vivante ☐ Hospitalisation ☐ Transfert vers un autre établissement ☐ Décès ☐ Sortie palliative ☐ Inconnu

DeepL: Résultat : ☐ Discharged alive ☐ Hospitalisation ☐ Transfert vers un autre établissement ☐ Décès ☐ Sortie palliative ☐ Inconnu

The extraordinarily high percentage of abbreviations and acronyms as causes for untranslated text (together accounting for no less than 69%) is, again, likely due to the fact that they are not sufficiently represented in the NMT's training data. This is underpinned by the observation that untranslated abbreviations and acronyms are often mistranslated in other segments of the text, leading to a combination of untranslated text, mistranslations, and inconsistent use of terminology (see Section 4.6). An instance where untranslated text is likely caused by the formatting is shown in ex. 10: heavy formatting, especially in cases where special characters are inserted, appears to impede a proper processing of the source text and consequently leads to a variety of issues, including untranslated text and issues related to consistency and grammar (see Section 4.5).

The severity of this error category mostly depends on the untranslated terms themselves as well as the context in which they appear. In this analysis, an equal distribution between MINOR (e.g. ex. 9) and MAJOR errors (e.g. ex. 8) could be observed. Major errors were most often annotated in relation to medical acronyms, as readers, including medical experts, may not be able to recognise or understand these terms in the source language. Untranslated text can hence lead to an impaired comprehensibility or even a complete loss of meaning.

4.5 Inconsistency

The category INCONSISTENCY captures all inconsistencies of non-terminological nature occurring within one text. Examples include the use of accepted orthographical variants (e.g. *plateforme* vs. *plate-forme*) as well as variations in grammatical register (formal vs. informal address of the reader). The investigation revealed that the most important cause of inconsistencies are enumerations (21%), followed by a heavy formatting of the source text (14%) and variants of grammatical gender (11%). The latter almost exclusively refers to the use of both the masculine and the feminine article for COVID-19. The relatively high percentage of inconsistencies caused by this factor can thus simply be explained by the nature of the texts analysed and is not necessarily transferable to other public health documents. The other two factors, however, are likely to be a frequent cause of inconsistencies regardless of the text genre:

(11) Inconsistency caused by an enumeration (139_45ff._G_ES)

ST: It is important to:
- Replace masks as soon as they become damp
- Dispose of masks immediately

REF: Es importante:
- Sustituir la mascarilla en cuanto se humedezca
- Desechar la mascarilla inmediatamente

Google: Es importante:
- Reemplace las máscaras tan pronto como se humedezcan
- Desechar las mascarillas inmediatamente

(12) Inconsistency caused by heavy formatting (145_46_G_ES)

ST: Sex at Birth ☐ Male ☐ Female ☐ Not specified

REF: Sexo al nacer ☐ Varón ☐ Mujer ☐ Sin especificar

Google: Sexo al nacer ☐ Masculino ☐ Mujer ☐ No especificado

It could be observed that enumerations such as in ex. 11 often lead to inconsistencies in combination with grammatical issues: given the preceding sentence, both verbs in the enumeration should be infinitives, but the first verb was translated as an imperative (reemplace) and the second as an infinitive (desechar). This is likely due to the disruption of the source text's continuity, which (similarly to heavy formatting) appears to impede a coherent processing of the immediate context. An inconsistent translation caused by the formatting is illustrated in ex. 12: in this case, the options related to the sex at birth were translated into Spanish with an adjective (masculino = male) as well as a noun (mujer = woman).

Non-terminological inconsistencies were almost exclusively classified as MINOR errors. However, similarly to issues related to style, such inconsistencies may still affect the acceptability and persuasiveness of a text. In some instances, they may also decrease its comprehensibility.

4.6 Combination of errors

Some of the examples above already suggested that there are many instances in which a single factor causes several different error types. Such a combination of errors, even if they are mostly minor ones, can have a considerable impact on the comprehensibility and acceptability of a text. This became particularly evident in the following example involving a medical acronym. Spanning over seven different segments, it had to be shortened for practical reasons and hence only shows the use of the acronym itself without providing the immediate context:[13]

(13) Combination of mistranslations, untranslated text, and inconsistent use of terminology caused by an acronym (164_6–99_D_FR)

ST: vaccine-preventable diseases (VPD) (seg. 6)
VPD (all other seg.)
REF: maladies à prévention vaccinale (MPV) (seg. 6)
MPV (all other seg.)

[13] This example was taken from a document that wasn't part of the pilot study, but has been analysed in the scope of the associated doctoral project. It is provided as a supplementary example here because it illustrates the considerable impact caused by a combination of errors particularly well.

DeepL: maladies évitables par la vaccination (MEV) (seg. 6)
MEV (seg. 7)
MVP (seg. 8)
SPV (seg. 81)
MPV (seg. 83)
VPD (seg. 98)
maladies transmissibles sexuellement (seg. 99)

In the English source text, the acronym VPD (referring to vaccine-preventable diseases) was first introduced with its full term and then used by itself throughout the rest of the document. The same was true for the human reference translation, but not for the machine-generated translations: DeepL started out with an adequate translation[14] of the full term and an introduction of a suitable acronym (seg. 6), which it used again in the next sentence (seg. 7). Just one sentence further, however, it mistranslated the term with a completely new and unrelated acronym (seg. 8). In the following paragraphs, the term was mistranslated again with another unrelated acronym (seg. 81), the acronym used by the human translator (seg. 83), and then left untranslated (seg. 98). In the very next sentence, DeepL dissolved the acronym and mistranslated it with "maladies transmissibles sexuellement" (sexually transmitted diseases). It is fairly obvious that this combination of mistranslations, untranslated text, and inconsistent use of terminology makes it virtually impossible for any reader, whether layman or expert, to follow and understand the text.

5 Conclusion

The analysis showed that current NMT systems are still prone to a variety of errors, ranging from minor stylistic issues to major mistranslations. Mistranslations were in fact the most frequent error type observed in this investigation, followed by inconsistent use of terminology, unidiomatic or awkward style, untranslated text, and other internal inconsistencies. The analysis of these errors vis-à-vis their causes revealed that ambiguity remains one of the most important challenges for NMT, often resulting in a complete change or loss of meaning. Other risk factors for major errors included technical terms, acronyms, and abbreviations, which can be particularly problematic for generic NMT systems due to their insufficient representation in the training data. The high incidence of issues related to terminological and textual consistency furthermore confirmed that current NMT systems are incapable of incorporating more than a few sentences as context, leading to difficulties with terminological variants, proper names, and enumerations. Lastly, it also became evident that the combination of errors caused by a single source can strongly amplify their negative impact.

These observations clearly show that raw NMT output is useful only to a very limited extent and that the risks and benefits associated with its use should be assessed extremely carefully. It is therefore crucial that decision-makers and users understand the limits and challenges of NMT and develop the ability to critically

[14] Albeit a violation of WHO's terminological preferences.

evaluate the quality and usefulness of its output. This is particularly important in safety-critical domains such as public health communication, where errors may have dramatic consequences, including financial damages, a loss of trust in authorities, and – most importantly – serious harm to people's health.

References

Almahasees, Zakaryia & Meqdadi, Samah & Albudairi, Yousef. 2021. Evaluation of Google Translate in rendering English Covid-19 texts into Arabic. *Journal of Language and Linguistic Studies* 17(4). 2065–2080. doi:10.52462/jlls.149.

Araabi, Ali & Monz, Christof & Niculae, Vlad. 2022. How effective is Byte Pair encoding for out-of-vocabulary words in Neural Machine Translation? In Duh, Kevin & Guzmán, Francisco (eds), *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas* (1). Association for Machine Translation in the Americas (Orlando, September 12-16, 2022). 117–130. doi:10.48550/arXiv.2208.05225.

Bowker, Lynne & Ciro, Jairo. 2019. *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Bingley: Emerald Publishing.

Haddow, Barry & Birch, Alexandra & Heafield, Kenneth. 2021. Machine translation in healthcare. In Susam-Saraeva, Şebnem & Spišiaková, Eva (eds), *The Routledge Handbook of Translation and Health*, 108–129. London: Routledge.

Isabelle, Pierre & Cherry, Colin & Foster, George. 2017. A challenge set approach to evaluating Machine Translation. In Palmer, Martha & Hwa, Rebecca & Riedel, Sebastian (eds), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen, September 7-11, 2017), Association for Computational Linguistics. 2486–2496. doi:10.18653/v1/D17-1263.

Khoong, Elaine & Steinbrook, Eric & Brown, Cortlyn & Fernandez, Alicia. 2019. Assessing the use of Google Translate for Spanish and Chinese translations of emergency Department Discharge Instructions. *JAMA Internal Medicine* (179). 580–582. doi:10.1001/jamainternmed.2018.7653.

Läubli, Samuel & Sennrich, Rico & Volk, Martin. 2018. Has Machine Translation achieved human parity? A case for document-level evaluation. In Riloff, Ellen & Chiang, David & Hockenmaier, Julia & Tsujiii, Jun'ichi (eds), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, October 31-November 1, 2018). Association for Computational Linguistics. 4791–4796. doi:10.18653/v1/D18-1512.

Lommel, Arle & Burchardt, Aljoscha & Uszkoreit, Hans. 2014. Multidimensional Quality Metrics (MQM): A framework for declaring and describing Translation Quality Metrics. *Tradumàtica: tecnologies de la traducció* (12). 455–463. doi:10.5565/rev/tradumatica.77.

Lommel, Arle & Görög, Attila & Melby, Alan & Uszkoreit, Hans & Burchardt, Aljoscha & Popović, Maja (2015): *QT21: Deliverable 3.1 – Harmonised Metric*. European Commission. doi:10.3030/645452.

Popović, Maja. 2018. Error classification and analysis for Machine Translation quality assessment. In Moorkens, Joss & Castilho, Sheila & Gaspari, Federico & Doherty, Stephen (eds), *Translation Quality Assessment*. Machine Translation: Technologies and Applications (1). Cham: Springer. doi:10.1007/978-3-319-91241-7_7.

Pym, Anthony & Ayvazyan, Nune & Prioleau, Jonathan. 2022. Should raw machine translation be used for public-health information? Suggestions for a multilingual communication policy in Catalonia. *Just. Journal of Language Rights & Minorities*, 1 (1-2). 71–99. doi:10.7203/Just.1.24880.

Šorak, Vanessa. 2025. Neural Machine Translation in pandemic-related health communication: A case study on risks and potentials in the context of the Covid-19 pandemic. In Atayan, Vahram & Choffat, Delphine & Czachur, Waldemar & Felder, Ekkehard & Pasques, Delphine (eds), *Diskursanalytische Perspektiven auf medizinische Fachkommunikation im europäischen Kontext*, 271–300. Heidelberg: Winter.

Vieira, Lucas Nunes & O'Hagan, Minako & O'Sullivan, Carol. 2020. Understanding the societal impacts of Machine Translation: A critical review of the literature on medical and legal use cases. *Information, Communication & Society* 24 (11). 1515–1532. doi:10.1080/1369118X.2020.1776370.

Zappatore, Marco & Ruggieri, Gilda. 2024. Adopting machine translation in the healthcare sector: A methodological multi-criteria review. *Computer Speech & Language* (84). 31–34. doi:10.1016/j.csl.2023.101582.