A Taxonomy and Study of Critical Errors in Machine Translation

Khetam Al Sharou[1] and **Lucia Specia**[1, 2]
[1]Language and Multimodal AI Lab, Imperial College London, UK
[2]Computer Science Department, University of Sheffield, UK
{k.al-sharou, l.specia}@imperial.ac.uk

Abstract

Not all machine mistranslations are of equal scale of severity. For example, mistranslating a date or time in an appointment, mistranslating a number or currency in a contract, or hallucinating profanity may lead to catastrophic consequences for the users. The severity of the errors is an important but overlooked aspect of machine translation (MT) quality evaluation. In this paper, we present the results of our effort to bring awareness to the problem of critical translation errors. We study, validate and extend an initial taxonomy of critical errors with the view of providing guidance for critical error analysis, annotation and mitigation. We test the extended taxonomy for three language pairs to examine to what extent it generalises across languages. We provide an account of factors that affect annotation tasks along with recommendations on how to improve annotation practice in future work. We also study patterns in the source text that can lead to critical errors. Detecting such linguistic patterns could be used to improve the performance of MT systems, especially for user-generated content.

1 Introduction

Machine Translation (MT) has now become ubiquitous in many online platforms (e.g. social networks) and generally used without any human post-editing due to cost, timeliness, and accessibility. The rapid development and adoption of MT has advanced efforts to improve and standardise MT evaluation, and increased discussion on how we should evaluate MT (Dorr et al., 2011; García, 2014; Ulitkin et al., 2021). This need escalated with the use of MT to translate user-generated content (UGC), e.g. in social media platforms. Unlike formal text, UGC often has colloquial language, including profanities, spelling errors, emojis, hashtags and abbreviations, and is grammatically ill-formed, which makes it hard for MT, often resulting in incorrect translations (Al Sharou et al., 2021). Some of these incorrect translations can contain critical errors. In this work, we refer to *critical errors* as instances of translations where the meaning in the target text deviates drastically from the source text where such translations can be misleading and may carry health, safety, legal, reputation, religious or financial implications.

The volume of content shared by users means that the MT-translated content cannot be manually post-edited. Therefore, users have to rely on MT as is and usually do not have the linguistic skills to identify the errors. As a consequence, users may be negatively affected if they misunderstand the intention or sentiment of the source text or could take inappropriate action if they act on critically corrupted translations. There are many instances where innocuous statements on social media have been translated by the machine to say something quite different, the opposite, or even turn a simple greeting into hate speech - translating 'good morning' in Arabic into 'attack them' in Hebrew by the machine, leading to the arrest of a Palestinian worker who posted it on his social media profile by Israeli police, as reported by the Guardian (Hern, 2017). Therefore, it is important that the issue of critical error is directly addressed.

To mitigate such a problem, recent research has

looked into automatic methods to detect critical errors in machine translation, with a view to inform users of such errors. This was framed as a track in the WMT 2021 Shared Task on Quality Estimation (Specia et al., 2021). A taxonomy was proposed to annotate training and evaluation data for this task. The annotation effort focused on critical errors only, i.e. other errors were disregarded. This differs from previous work, where critical errors – if evaluated – are seen as an extra level of annotation on general errors, i.e. as a *severity judgement* on errors (Lommel et al., 2014). From a practical perspective, we believe this focused annotation is a good strategy as it saves annotation effort and allows gisting-oriented quality prediction models, under the assumption that MT is still usable even though it may contain minor (non-critical) errors. According to Specia et al. (2021), however, the annotation of critical errors proved very challenging, with low agreement amongst annotators.

A taxonomy is an important step as it establishes which types of errors should be considered critical. We revisit and extend the taxonomy proposed in Specia et al. (2021) in order to (a) perform a more focused, smaller-scale study with well-trained annotators to understand the general challenges in annotating critical errors, and (b) validate the extended taxonomy on different languages. For that, we commission the manual annotation of such errors and conduct an in-depth analysis of their impact on the translations. We reflect on the annotation process as an essential part of any evaluation task that aims to examine the performance and usability of MT systems for better evaluation and annotation practices. We also show how the source text can affect the quality of MT translations when it comes to the presence of critical errors.

We start by presenting an overview of popular quality evaluation taxonomies (Section 2) to then introduce the taxonomy we study, developed in Specia et al. (2021), with two additional categories we propose to add to the taxonomy (Section 3). We then explain our approach and criteria to validating the extended taxonomy and follow that with a data analysis through which we show how the taxonomy is validated (Section 4). We also reflect on the annotation process for different languages (Section 5). Finally, we explore how the quality or lack of quality of the source text could contribute to the generation of critical errors (Section 6).

2 Related Work

With the rapid development and increasing adoption of Machine Translation systems, evaluating the quality has become a common practice. This has led to advances in the area of translation quality assessment (TQA) and inspired initiatives that aimed to standardise this practice.[1] TQA is used to assess the performance of a system, and whether its output fits to be used either as is or as a first draft that requires some post-editing (O'Brien, 2012; Han, 2022). TQA can also be utilised to enhance the performance of systems, as a point of comparison between various systems, or to estimate the effort required to post-edit machine-translated content (Aziz et al., 2012; Popović, 2018). Examining the quality of the MT output has been conducted through either the identification of errors, the overall assessment of MT quality or both.

Various classifications of errors have been developed, against which MT system outputs are assessed (Lommel et al., 2014; Abu-Ayyash, 2017; Popović, 2018). The two most comprehensive frameworks, which have been widely adopted in industry, academia and by end-users, are (i) Multidimensional Quality Metrics (MQM), proposed under the EU-funded QTLaunchPad project (Lommel et al., 2014), and (ii) Dynamic Quality Framework (DQF) by the Translation Automation User Society (TAUS) (Lommel et al., 2015; Rivera-Trigueros, 2021). These initiatives offering general taxonomies are based on, and inspired by, earlier error-specific models including LISA QA Model, developed in the 1990s by the Localisation Industry Standards, and the SAE J2450 metric, among others (Lommel et al., 2014).

Another group of individual error classifications includes language-related and linguistically-motivated taxonomies that aim to evaluate the quality of MT output according to specific linguistic phenomena that occur in the translation and are associated with certain languages. For example, Costa et al. (2015)'s study classifies translation errors from English into European Portuguese. Their work extends previous taxonomies to study errors associated with morphologically rich languages. Some other studies focus specifically on the impact of certain features of the text on the output. For example, Abu-Ayyash (2017) explores errors and non-errors for the English-Arabic pair in MT-translated gender-bound constructs in tech-

[1] In this work, we only focus on human evaluation.

nical texts, and Han et al. (2020) proposes a categorisation of error types generated by MT systems when translating multiword expressions.

In addition to classifying types of errors, other aspects of quality evaluation are considered, i.e. the importance and severity of the errors. Still, these are optional criteria and considered depending on the task and the purpose of the translation. In the MQM framework, importance is assigned to categories of errors. For example, if one category is considered as a priority for a given task, it is deemed as important for that specific task. Severity, however, is applicable to individual errors, and is related to their nature and their impact on the usability of the translation. 'The more severe an error is, the more likely it is to negatively affect the user in some fashion' (Lommel, 2018). MQM identifies four levels of severity: critical, major, minor, and null that align to some extent with those adopted in the DQF framework (Lommel, 2018).

More recent work has focused on classifying only the most severe errors (referred to as *critical errors*). For example, the WMT 2021 Shared Task on Quality Estimation (Specia et al., 2021) organised a track on predicting the presence of critical errors in sentence translations. As part of this track, a taxonomy of critical errors was proposed and a large amount of data was annotated for such errors: 10K translations from English into four languages (Chinese, Japanese, Czech and German). Each translation was annotated by three professional translators. However, the authors observed that the annotation was problematic, with overall low annotator agreement. It was not clear from the effort whether this was because of the general lack of understanding of the task by the annotators, the complexity of the task or because of other factors.

One interesting outcome of the report in Specia et al. (2021) was the high proportion of critical errors in UGC. It is clear that error-free MT is still unattainable and that critical errors are not rare. Therefore, further research towards understanding, formalising, and annotating such errors is much needed before prediction and mitigation strategies can be put in place. We, therefore, devote this work to bring attention to this issue. We study critical errors that have the same level of severity (highest), and treat them as critical errors because of their potential negative impact on those who use the translations as they are. The assumption, which we test in this paper, is that the types of critical errors should be applicable to any language pair. As far as we know, this is the first work which focuses on studying critical errors in UGC.

3 A Taxonomy of Critical Errors

In what follows, we present Specia et al. (2021)'s taxonomy of critical errors a) to serve as the base for a new extended taxonomy developed in this work and b) to be tested and analysed in detail. It recognises three ways in which meaning deviations from the source sentence can happen:

- **Mistranslation**: content is translated incorrectly into a different meaning, copied to the target text (i.e. it remains in the source language), or translated into gibberish.
- **Hallucination**: content that is not in the source is introduced into the translation. For example, profanity words are introduced.
- **Deletion**: critical content that is in the source sentence is not present in the translation. For instance, the source sentence may contain a negation that is removed from the translation.

In this taxonomy, there are five main categories of critical errors:

1. Deviation in toxicity (TOX): This category refers to instances where the translation may incite hate, violence, profanity or abuse against an individual or a group (a religion, race, gender, etc.) due to incorrect translations. It covers cases where toxicity is introduced into the translation when it is not in the source, deleted in the translation when it is in the source, mistranslated into different (toxic or not) words, or not translated at all (i.e. the toxicity remains in the source language or transliterated).

2. Deviation in health/safety risks (SAF): This category refers to instances where the translation may bring a risk to the reader where the meaning which has been changed has health and safety implications. This issue can happen when content is introduced into the translation, deleted from the translation when it is in the source, or mistranslated into different words, or not translated at all (i.e. it remains in the source language).

3. Deviation in named entities (NAM): A named entity (people, organisation, location) is deleted, mistranslated by either another incorrect named entity or a common word or gibberish, left untranslated when it should be translated, or introduced when it is not in the source text.

4. Deviation in sentiment or negation (SEN):

The MT either introduces or removes a negation (with or without an explicit negation word), or reverses the sentiment of the sentence (e.g. a negative sentence becomes positive or vice-versa).

5. Deviation in numbers, time, units, or date (NUM): The MT mistranslates or removes a number, date, time or unit, causing misunderstanding that could lead to an unpleasant, or major, consequence such as missing an important appointment. In this work, we propose two additional categories to add to the taxonomy:

6. Deviation in instructions (INS): This category refers to instances where the MT translates instructions incorrectly, such that if one were to follow them, they would not get to the intended outcome (except for negation and reversal of sentiment cases - category SEN). This also includes cases where pronouns are changed.

7. Other critical meaning deviation (OTH) - specify: This category involves instances of translations where the meaning changes in a critical way which does not come under any of the abovementioned categories. For example, the MT system could change the meaning of a verb or a phrase completely or distort the structure of a sentence, affecting its intended meaning, e.g. by locating the object of the sentence in the place of the subject.

4 Validating the Taxonomy

In this section, we report on a study we performed on this extended taxonomy by means of an annotation exercise with additional languages, followed by an in-depth analysis.

4.1 Data Annotation

We have carried out the annotation process to validate the extended taxonomy as follows: We have manually selected 100 sentences (roughly 2000 words) from the WMT21 Critical Error Detection task dataset. The original English data comes from the Wikipedia Comments Corpus.[2] Our selection was motivated and based on Al Sharou et al. (2021)'s work on non-standard text and used their categories of non-standard linguistic features that can be challenging to the machine. Based on that, half of the sentences selected included features such as abbreviations, special characters, spelling mistakes, wrong punctuation marks, among others. We also chose sentences that contained offensive language, which the MT system is less likely to have been trained to handle it. The expectation is that such sentences may lead to critical errors when translated automatically. The other half did not include any of the said features. We targeted three language pairs, i.e. English–Chinese, English–Italian and English–Arabic, to test whether the extended taxonomy of critical errors is applicable to languages from different families. Still, this annotation task is not merely about other languages, but it is also a more focused effort, carried out with trained annotators. To translate the data, we used three MT systems, Google Translate, Bing and Systran.[3] The initial data in Specia et al. (2021) only used one translation system, i.e. the ML50 fairseq multilingual Transformer model (Tang et al., 2020)[4]. For each language, we asked three translators who are native speakers of these languages to carry out the manual annotation. Their professional translation experience ranges from two to six years, and two of them have experience carrying out annotation tasks. Annotators were provided with the extended taxonomy of critical errors. Online sessions were held with them to explain the purpose of the study along with the extended taxonomy and followed up by email communications to solve any issues they had encountered while carrying out the task. Annotators were provided with clear guidelines where they had to strictly follow two main rules:

- *This evaluation is NOT about flagging any mistranslation/hallucination/deletion errors, but only cases where such errors are critical and lead to catastrophic consequences, as outlined in the Taxonomy of Errors.*
- *This evaluation is NOT about flagging toxicity (hate, profanity) in the translation, but rather cases where the meaning in the translation differs from the content in the source in a critical way.*

We asked annotators to label the data at the sentence-level with a binary label, where the occurrence of one or more errors means the sentence has critical errors. We also requested them to assign the type of error, selected from a drop-down list, based on the extended taxonomy, to the first critical error they find. We used multiple annotators to measure agreement levels as one of our met-

[2] https://meta.wikimedia.org/wiki/Research:Detox

[3] The online systems were used between November 2021 and March 2022.

[4] https://github.com/pytorch/fairseq/tree/master/examples/multilingual

rics to validate the taxonomy and annotation task. Given the small number of participants, which may undermine the effectiveness of statistical analysis, we also look at the results from a qualitative perspective. We also asked the annotators to complete a questionnaire, reflecting on their experience carrying out the annotation task. The annotators were instructed to conduct the annotation independently.

4.2 Data Analysis

In order to validate the extended taxonomy, we looked at the annotation carried out for the three languages in light of two criteria:

- **Reproducibility** (through agreement rate among annotators): by confirming the presence or absence of critical errors in each translation, regardless of the types of critical error(s).
- **Applicability to other languages**: whether the error types in the taxonomy are observed for different language pairs.

4.2.1 Reproducibility

In this section, we present an analysis of the inter-annotator agreement (IAA) ratings among annotators, based on the set of 100 sentences, for each of the three language combinations, i.e. English–Chinese (EN–ZH), English–Italian (EN–IT) and English–Arabic (EN–AR).

Sentence Level: We compute IAA on the sentence-level binary labels, using Cohen's Kappa (Cohen, 1960), where raters agree on whether or not the sentence has at least one critical error, regardless of the type of critical error.

Table 1 displays the results for error mark-up, presented in pair-wise comparisons to evaluate the similarity between each pair of annotators.

Annot.	EN–ZH	EN–IT	EN–AR
1&2	0.802	0.906	0.840
2&3	0.825	0.652	0.640
1&3	0.872	0.699	0.640
Average	0.833	0.752	0.706

Table 1: Cohen's Kappa IAA - Sentence Level

Table 1 shows a substantial agreement among the annotators across the three languages, with English–Arabic gaining the lowest agreement rating. This high rating could have been influenced by the way the dataset was selected, described in Section (4). It is of relevance to note that although Arabic annotators (**2&3**) and (**1&3**) have the same agreement rating, their rating shows some discrepancies when it comes to error types (see Table 2) below. It is also important to clarify that we intended to order annotators according to whether they had received training on the taxonomy and guidelines (annotators labelled as **1**), followed by those who did not attend but asked for clarification (annotators labelled as **2**), then the ones who carried out the annotation using only the guidelines and the extended taxonomy (annotators labelled as **3**). This explains why the agreement rate among annotators (**1&2**) is higher, especially for the English–Italian and English–Arabic language pairs. These results serve our aim to examine factors such as training that can affect the annotation task and annotators' performance (for an in-depth analysis of the annotation task, see Section 5).

Type Level: As a further step, we calculate the IAA on a categorical scale. We use Fleiss' kappa in SPSS (Fleiss, 1971; Fleiss et al., 2003) that allows determining the level of agreement on a categorical scale, i.e. agreement on individual categories of errors. Based on the extended taxonomy, we included in the annotation task, as a dropdown menu for the annotators to use, the seven categories in addition to one more category, labelled as 'None', to cover cases where no critical error(s) were detected. Results presented in Table 2 show that the average over all pairs of annotators and all categories is lower in all languages, compared with the sentence level agreement rating. Overall categorical agreement rating can be described as moderate for Italian and Arabic (**0.548** and **0.424** respectively), and substantial for Chinese (**0.624**). This reveals that annotators may have found it difficult to decide on the types of errors. Their assessment may have been influenced by several factors. Annotation is to some extent a subjective task and is greatly influenced by how annotators treat and understand the source and target sides of the data. For example, some annotators were inclined to label errors as critical based on their own assessment rather than according to what the guidelines say (see discussion in Section 5).

It is interesting to see that Chinese has the highest agreement rate in both rating exercises, i.e. sentence level (**0.833**) and type level (**0.624**). A closer look shows that error types were assigned mainly under three types, i.e. '**TOX**', '**Other**' and '**None**'. This somehow explains why it has the highest average agreement rates at both levels. We also no-

Error Type	Annot.	EN–ZH	EN–IT	EN–AR
TOX	1&2	0.451	0.792	0.838
	2&3	0.968	0.452	0.552
	1&3	0.336	0.435	0.535
SAF	1&2	—	-0.005	—
	2&3	-0.005	1	—
	1&3	-0.005	-0.005	—
NAM	1&2	-0.005	-0.02	-0.015
	2&3	—	0.490	-0.005
	1&3	-0.005	-0.01	-0.01
SEN	1&2	—	—	-0.01
	2&3	—	-0.005	-0.005
	1&3	—	-0.005	-0.005
NUM	1&2	—	-0.005	—
	2&3	—	-0.005	—
	1&3	—	1	—
INS	1&2	0.011	-0.015	-0.015
	2&3	0.795	-0.01	0.096
	1&3	0	0.385	-0.111
Other	1&2	0.479	-0.005	-0.031
	2&3	0.740	-0.02	—
	1&3	0.656	-0.026	-0.031
None	1&2	0.757	0.906	0.640
	2&3	0.872	0.486	0.880
	1&3	0.944	0.532	0.640
Overall Agreement		0.624	0.548	0.424

Table 2: Fleiss' kappa Agreement on Error Types

tice that annotators (**2&3**) are closer in their agreement rates, especially when it comes to '**Other**' and '**None**' categories. These two annotators may have collaborated on this task, although annotators were told to work independently.

It is important to highlight that a high rate is given to certain categories, e.g. '**INS**' in Chinese, achieving **0.795**. When annotators (**2&3**) from this group asked about the reason behind their selection of the '**INS**' type, their answer showed that they interpreted sentences in the *seemingly* imperative format as instructions, hence assigned errors as a 'deviation in instructions'. In reality, this might not have been the case, especially that the Chinese annotator **1** and the annotators for the other language pairs did not label a similar number of critical errors under the '**INS**' category. This finding gives an indication about how failure to understand what each category implies by annotators could affect the evaluation and annotation task and necessitates that focused training is provided, especially when more specific tasks are assigned to annotators.

4.2.2 Applicability

We carried out an analysis to validate the applicability of the extended taxonomy. Namely, we

- present an analysis of the error distribution for the language pairs, i.e. English–Chinese, English–Italian and English–Arabic;
- provide examples of the different types of errors in the three selected language pairs.

Error distribution across the three languages is presented in Table 3. We calculate the average of the total number of each error type, selected by the three annotators, for each language pair to show how many times each error has been selected by the annotators across the three languages.

Annot.	EN–ZH	EN–IT	EN–AR
TOX	16.33	24.33	38.7
SAF	0.33	0.67	—
NAM	0.33	1.67	1
SEN	—	0.33	0.67
NUM	—	0.67	—
INS	7.33	1.67	7.33
Other	3.67	1.67	2
None	72.00	69.00	50.3

Table 3: Error Distribution across Languages

The majority of types in the extended taxonomy have occurred in the dataset analysed for the three language pairs. In a few cases, some types did not occur at all as in the Chinese side of the dataset, i.e. '**SEN**' and '**NUM**', and the Arabic side, i.e. '**SAF**' and '**NUM**'. The two types with the highest number of occurrences are '**TOX**' and '**None**', albeit with different proportions. The occurrence of '**TOX**' type could be as a result of the type of the annotated data which has a substantial amount of offensive language. This aspect of the text, when existing in large quantities, could lead to the generation of critical errors. '**None**' type is the most selected type among the types across the three languages. This could be attributed to the fact that half of the dataset (50 sentences) did not include features that are challenging to the machine (e.g. no offensive language or non-standard features, hence, less causes of critical errors). This finding shows the impact of the source text on the output. We expand on this aspect extensively in a separate Section (6), due to its importance in affecting online communication and also for consideration by any future work that aims to improve the quality of MT systems and develop error and noise analysis and detection models. Some types such as '**NUM**' did not appear much as the sentences did not have information that could lead to errors of this type. These findings prove that the types included in the extended taxonomy can occur in different languages. This also shows that MT systems behave differently depending on the language. For

example, while the annotators did not find errors that fall under '**SAF**' and '**NUM**' categories in the Arabic side of the dataset, and under '**SEN**' and '**NUM**' in the Chinese, that was not the case in the Italian side of the dataset which covered all types of errors.

Examples provided were chosen as an illustration for their clarity and strong manifestation of deviation to show how far the machine can go in generating critical errors when translating UGC. These examples were obtained from the analysis of the dataset, covering the three chosen languages. The examples with their translations are provided in English only, following the order of the types in the extended taxonomy (see Section 3).

Deviation in toxicity (TOX)

ST	**Your killing the fucking planet**.
MT-ed text	**May the damn planet kill you**.
Translation into Arabic by Systran	

Deviation in health/safety risks (SAF)

ST	I Know two teenagers that suffer from **gerd** it is a big problem for these people!
MT-ed text	I Know two teenagers that suffer from **root disease** it is a big problem for these people!
Translation into Chinese by GT	

Deviation in named entities (NAM)

ST	Your fucking ass doesn't know shit about it AT ALL.**Rocky**.
MT-ed text	Your fucking ass doesn't know shit about it AT ALL.**rock**.
Translation into Italian by Bing	

Deviation in sentiment or negation (SEN)

ST	Don't the Yoshinoyasin Singapore and Indonesia ALSO **not serve** pork?
MT-ed text	Don't the Yoshinoyasin in Singapore and Indonesia ALSO **serve** pork?
Translation into Arabic by GT	

Deviation in numbers/time/units/date (NUM)

ST	Your signature is incredibly long. At 632 characters, it's about **two and a half times** what the software allows.
MT-ed text	Your signature is incredibly long. At 632 characters, it is **double** what the software allows.
Translation into Arabic by GT	

Deviation in instructions (INS)

ST	The link to wikibooks doesn't work and I don't know how to fix it. **Can anyone help?**
MT-ed text	The link to wikibooks doesn't work and I don't know how to fix it. **Can I help you?**
Translation into Arabic by GT	

Other critical meaning deviation (OTH)

ST	Admin's **beware** of him.
MT-ed text	Admin is **aware** of him.
Translation into Italian by Systran	

As a further step in our effort to validate the taxonomy, we reflect on the annotation process, using data collected through post-annotation questionnaires and our own experience supervising the annotation process. We also look at the impact of the source text on the generation of critical errors.

5 Evaluation of the Annotation Task: Challenges and Recommendations

Data was annotated for the three selected languages by professional translators. We provided them with guidelines based on the extended taxonomy with clear instructions that they must only annotate critical errors with catastrophic impact on the translation. However, we have found that:

- Despite providing clear guidelines on critical errors and how to detect and categorise them, there was some disagreement among the annotators regarding what errors were considered critical. This led them to tagging errors as critical when they were not, and vice versa.
- Annotators found it difficult to focus on critical errors versus annotating all errors.

These findings pose the following questions: (1) how this task is conducted?, 2) what areas need to be addressed for the annotation to be carried out at a level that serves the purpose of the annotation task?, and (3) what makes annotating critical errors a difficult task? We reflect on these areas and present a set of factors along with recommendations, based on empirical findings, with the aim to improve the annotation process for future work.

- **Training**: Training is important to ensure annotators understand the task. The role of training is displayed in the differences in the annotation between those who joined the training and those who only followed the guidelines without training. A follow-up discussion with the second group whose annotation contained major differences revealed that there was some misunderstanding regarding what each category implied, failing to analyse the translations correctly as a result.
- **Difficulty and specificity of the task**: Disagreement among annotators occurred because the task was not easy for them. To clar-

ify further, some annotators found it difficult to just focus on critical errors and disregard other errors as a new practice they have not experienced before. This finding highlights that general training might not be enough to understand the requirements of more specific annotation tasks.

- **Prior attitude towards the annotation task**: Some annotators felt unsure about why such translations with critical errors should be accepted and the purpose of carrying out the annotation task. These annotators tended to consider errors as critical when they did not follow the general rules of a language (grammatical or stylistic rules), overlooking what the guidelines stated, ending up annotating both minor and critical errors. It is, therefore, vital to not only provide clear instructions on how to carry out the annotation task, but to also highlight that they need to treat it as a serious task similar to translating official documents and that they should always follow the guidelines (i.e. annotation brief).
- **Time allocated to the task**: Annotators were involved on a voluntary basis which could have limited the time they allocated to performing the annotation task. Annotators reported spending between 2-8 hours on this task. Therefore, annotators who spent less time might not have worked on it thoroughly, affecting the quality of their annotation.
- **Subjectivity of the task**: Although clear guidelines were provided, annotators differed in their interpretation of each type. Their understanding of the translations also affected their judgement of whether the errors were critical or not. Where disagreements occurred, we asked them to provide their interpretation of the source text and the translations and the reasons which influenced their decision. This helped us understand whether the guidelines or their understanding of the translation contributed to the disagreement.
- **Communication with annotators**: Some annotators were hesitant to ask for clarification, fearing that might show them as less qualified. It is, therefore, vital to establishing communication with annotators while conducting the annotation task for a better performance.
- **Misleading translations**: Some instances of disagreement occurred as the annotators only read the translations without referring back to the source text. This happened where the translation sounded fluent in the target language. This finding highlights the need to consider both source and target texts to determine whether an error is critical.

6 Source-text Impact on the MT Output

This section presents an analysis of the source text to show whether there is a correlation between the quality of the source text and the generation of critical errors. For this purpose, we analyse translations produced by the three online MT systems (Google Translate, Bing and Systran) for one language combination, i.e. English-Arabic, using the same dataset (100 sentences). Our focus on Arabic was driven by the availability of language expertise (i.e. one of the authors is a native speaker of Arabic). The assumption is that if the different systems struggle with the same source sentences, producing critical errors, it would give indications about the potential output the machine could produce when handling such texts. Our aim is to detect patterns in source sentences that can cause critical errors to be considered when developing MT systems to improve the performance of such systems, especially for UGC. We use, as a point of reference, Al Sharou et al. (2021)'s taxonomy of aspects of non-standard text that could affect the quality of the translation. For readability, back translations of the errors are provided in English.

Offensive language The importance of looking at this aspect of the data comes from its extensive existence in UGC and its sever impact on the output. Our analysis shows that most translations that have critical errors are those of sentences which contain offensive language. When the sentence has a large number of swearing/offensive words and idiomatic phrases, the machine tends to produce wrong translations that are unreadable or completely different from the source. When it comes to translating offensive language, we recognise the use of different 'strategies' including literal translation, transliteration, omission, random translation (hallucination) or substitution of one strong word with another milder word and vice versa. Sometimes, the machine uses a mix of these strategies when translating the same sentence, failing to convey correct translations as a result. For example, the three systems failed to provide correct translations of the offensive language in this

sentence '***Piss off Homo**, no one wants to hear from you, also hahahahaha **you can't get married #asshole***', leading to major errors which have affected the original meaning. These systems vary in how they handled this type of language. GT translated '**piss off**' as '**rape**', while Bing ignored '**off**' as being part of the verb and translated '**piss**' as '**urinate**' and '**off**' as '**in front off**'. '**Homo**' was transliterated by both GT and Bing, and Systran mistranslated it as '**human**', affecting the meaning of the last part of the sentence '**you can't get married**', which was deleted by GT but reserved by Bing and Systran. The swearing word '**asshole**' was left untranslated by Bing and Systran and deleted by GT.

Symbols and special characters The use of special symbols/characters such as star signs (*) or hashtags (#) can lead to erroneous translations. MT tends to overlook words which contain such special characters, render incorrect meaning or leave it in its source language. Arabic translation of the words that have been disguised by replacing letters with star signs in the sentence '*Stop being such an **a**hole...you **f***ing re***d***' shows that the three systems have either preserved the star signs and translated what left as another word, e.g. rendering '**a**hole**' as '**hole**' with the two star signs coming after it, or preserving it as random letters, conveying no meaning, as in the translation of '**f***ing re***d**' by Bing as '**g**' '**d**'; or dropped completely by GT and Systran.

Punctuation marks Misusing punctuation marks (e.g. deletion, addition, or use of wrong punctuation marks), especially commas and full stops, could lead to a mix up of the different parts of the sentence or different sentences, generating critical errors. For example, the translation of '***I give up Thanks** for ruining the Lion King pages*' shows the impact a missing punctuation mark has on the translation. The three systems translated the first part as '**I gave up thanking**'. They, therefore, do not deliver the original meaning where the writer intended to say he/she is giving up trying to keep the pages, and that the word '**thanks**' is used in a sarcastic way to express his/her frustration.

Negation Negation can lead to critical errors when reversed from negative to positive or vice versa; through e.g. dropping or reversing negative words (e.g. not, never, nobody); or reversing the meaning of some words (for instance, the three systems translated the verb **'reverting'** in '*why keep **reverting** my edits?*' as **'bringing back'**.

Named entities Named entities can be confusing to the machine especially when the name has different meanings and the MT system fails to treat it as a proper name, or when the names are unknown to the machine. Names are either mistranslated, left untranslated or deleted completely. For instance, Bing translated the proper name '**Rocky**' in the sentence '*your fucking ass doesn't know shit about it AT ALL.**Rocky***' as a noun rather than transliterating it, resulting in a wrong translation, while GT and Systran dropped it completely.

Spelling mistakes and contractions When dealing with spelling mistakes and informal contractions, the machine gives a translation that does not reflect what the source text says. In other cases, the machine preserves them in their original language or transliterates them. For example, the word '**freakin**' is transliterated by GT and left untranslated in the translations provided by Bing and Systran when translating this sentence '*Dude, u got a stick in ur ass, **lemme** edit the **freakin** montana academy page!*'. The short form of '**let me**' '**lemme**' is left untranslated by Bing while transliterated by GT and Systran, making it sound like a proper name where the translation in Arabic reads as '***Lemme edited montana academy page***'.

Capital letters Random capitalisation seems to affect the MT output. The analysis of the dataset shows that the three systems treated words written in capital letters as proper names. For example, the linking verb **'IS'** in the sentence *'The fact is 'Irish' is the commonly used term in Ireland and Wiki seeks to reflect what **IS** rather than what might be correct'* was translated by the three systems as '**Islamic State (or Daesh)**'. Such a translation could pose a potential risk if it were actually used in a sensitive context.

Lack of pronouns The lack of pronouns can lead to critical errors where the machine randomly replaces one pronoun with another. In this example, *'didn't forget, just been busy - will find the time to look into it'*, '**didn't forget**' was translated as '**don't forget**' by GT, '**he didn't forget**' by Bing, and only correctly translated by Systran as '**I didn't forget**'; '**just been busy**' was translated as '**I was busy**', '**he was busy**' and '**I was busy**' respectively. The three systems wrongly rendered

'**will find the time**' as '**you will find the time**'.

7 Conclusion

This work validated an extended taxonomy of critical errors developed to serve as a stand-alone taxonomy that can be used to evaluate or detect critical errors in machine-translated content. Findings emphasise the need to address critical errors with catastrophic impact on the output and for further attention to be paid not only to developing guidelines on critical errors, but to also training annotators on how to spot and assess them. It has proved that critical errors are not rare, and they are not specific to certain languages. It also underlines the need to improve current MT systems to specifically deal with user-generated content, considering aspects of the text that could lead to critical errors to improve online communication and enhance MT's role in enabling, rather than hindering, communication among speakers of different languages.

Acknowledgement

Lucia Specia was supported by funding from the Bergamot project (EU H2020 Grant No. 825303).

References

Abu-Ayyash, Emad AS. 2017. Errors and non-errors in english-arabic machine translation of gender-bound constructs in technical texts. *Procedia Computer Science*, 117:73–80.

Al Sharou, Khetam, Zhenhao Li, and Lucia Specia. 2021. Towards a better understanding of noise in natural language processing. In *RANLP 2021*, pages 53–62.

Aziz, Wilker, Sheila Castilho, and Lucia Specia. 2012. Pet: a tool for post-editing and assessing machine translation. In *LREC-2012*, pages 3982–3987.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Costa, Ângela, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur. 2015. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 29(2):127–161.

Dorr, Bonnie, Joseph Olive, John McCary, and Caitlin Christianson. 2011. Machine translation evaluation and optimization. In *Handbook of natural language processing and machine translation*, pages 745–843. Springer.

Fleiss, Joseph L, Bruce Levin, and Myunghee Cho Paik. 2003. *Statistical methods for rates and proportions*. john wiley & sons.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

García, Ignacio. 2014. Training quality evaluators. *Revista Tradumàtica: tecnologies de la traducció*, (12):430–436.

Han, Lifeng, Gareth Jones, and Alan Smeaton. 2020. Alphamwe: Construction of multilingual parallel corpora with mwe annotations. *arXiv preprint arXiv:2011.03783*.

Han, Lifeng. 2022. An overview on machine translation evaluation. *arXiv preprint arXiv:2202.11027*.

Hern, Alex. 2017. Facebook translates' good morning'into'attack them', leading to arrest. *the Guardian*, 24.

Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció*, (12):455–463.

Lommel, Arle, Attila Görög, Alan Melby, Hans Uszkoreit, Aljoscha Burchardt, and Maja Popović. 2015. Harmonised metric. *Project Report, QT21 Project*.

Lommel, Arle. 2018. Metrics for translation quality assessment: a case for standardising error typologies. In *Translation Quality Assessment*, pages 109–127. Springer.

O'Brien, Sharon. 2012. Towards a dynamic quality evaluation model for translation. *The Journal of Specialised Translation*, 17(1):55–77.

Popović, Maja. 2018. Error classification and analysis for machine translation quality assessment. In *Translation quality assessment*, pages 129–158. Springer.

Rivera-Trigueros, Irene. 2021. Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, pages 1–27.

Specia, Lucia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André Martins. 2021. Findings of the wmt 2021 shared task on quality estimation. Association for Computational Linguistics.

Tang, Y., C. Tran, Xian Li, P. Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401.

Ulitkin, Ilya, Irina Filippova, Natalia Ivanova, and Alexey Poroykov. 2021. Automatic evaluation of the quality of machine translation of a scientific text: the results of a five-year-long experiment. In *E3S Web of Conferences*.