Agent and User-Generated Content and its Impact on Customer Support MT

Madalena Gonçalves[*] **Marianna Buchicchio**[*] **Craig Stewart**[*]
Helena Moniz[*†‡] **Alon Lavie**[*]
[*]Unbabel
[†]INESC-ID, Lisboa, Portugal
[‡]University of Lisbon, Lisboa, Portugal
[*]{firstname.lastname}@unbabel.com

Abstract

This paper illustrates a new evaluation framework developed at Unbabel for measuring the quality of source language text and its effect on both Machine Translation (MT) and Human Post-Edition (PE) performed by non-professional post-editors. We examine both agent and user-generated content from the Customer Support domain and propose that differentiating the two is crucial to obtaining high quality translation output. Furthermore, we present results of initial experimentation with a new evaluation typology based on the Multidimensional Quality Metrics (MQM) Framework (Lommel et al., 2014), specifically tailored toward the evaluation of source language text. We show how the MQM Framework (Lommel et al., 2014) can be adapted to assess errors of monolingual source texts and demonstrate how very specific source errors propagate to the MT and PE targets. Finally, we illustrate how MT systems are not robust enough to handle specific types of source noise in the context of Customer Support data.

1 Introduction

Unbabel's Language Operations platform blends advanced artificial intelligence with humans in the loop, for fast, efficient, high-quality translations that get smarter over time. The company combines Machine Translation with Human Post-Edition performed by non-professional post-editors to translate Customer Support content in a variety of formats including emails and chat messages. Customer Support is a highly unique domain given that it involves bilateral communication between Customer Support agents ('agents') and customers ('users'), each with their own nuanced discourse strategies and features.

Notwithstanding, primary literature such as Nars et al. (2016), generally consider both sides of the interaction jointly, without regard for independent and differentiating factors. The fundamental differences can be characterized as follows: agents are call center employees who are usually non-native speakers of English, which has generally evolved as the 'lingua franca' of Customer Support and the primary source language translated at Unbabel. Given that English is usually not the agent's first language, it is common to observe elements of language transfer (where the grammar rules of their native language are transferred to the English language) and other linguistic errors, more commonly the addition and omission of prepositions and articles, as also mentioned in Lee and Seneff (2008) and Rozovskaya and Roth (2010). According to Sinha et al. (2009), a person's experience and knowledge of their mother tongue will most definitely interfere with the learning of a second language, thus creating errors of different nature.

The interaction established by agents is somewhat controlled because they are usually following a particular protocol for communication prescribed by their company. This might include response templates, branding and fixed terminology which discourages stylistic variance and can often result in a large amount of repetition both within and across interactions.

Agents also work quickly, aiming to provide consistently timely responses. This can often result in typographical and other linguistic errors.

Agents often operate in highly stressful circumstances; they might have to meet certain quotas which, for example, demand a brief turnaround time. This will ultimately influence their performance and the introduction of errors in the messages which, in our use case, are subsequently translated by MT and post-edited.

User-generated content from customers, on the other hand, is highly variable and unstructured. Common features include the use of abbreviations, emoticons and idiomatic expressions, all of which present unique challenges to MT. Grammatical and typographical errors common to user-generated content resulting from keyboard or smartphone use are also present.

Most critically, content from customers in a Customer Service context is highly purposeful and sometimes emotionally volatile. Customers often contact customer support to complain about a product or service and may exhibit high levels of impatience and frustration. Linguistically, this is reflected in unique lexical choices, such as the use of profanities, and variable capitalization and punctuation. All of which can often result in degradation of translation quality where the interaction is translated into or out of the source language.

Additionally, the native language of the customer is not always predictable. As mentioned in Roturier and Bensadoun (2011), often times customers from non-English speaking countries will be interacting in a non-native tongue. This could be either because they live in a foreign country, or because they are engaging the services of a foreign company. Finally, as mentioned in Hohn et al. (2016), being a native speaker of a language does not directly indicate a high proficiency of that language, another factor that can potentially indicate poor source text inputs.

Different types of Customer Support content also present another dimension of complexity: consider, for example, how a chat might differ from an email. The response time required in the former will often determine the fidelity and quality of the resulting interaction. As Lind (2012) illustrates, time restrictions implied in chat language ultimately result in fragmented written content. Because emails do not require real-time translation, at Unbabel, translation is performed first by MT and subsequently post-edited. Chat messages, however, require instantaneous translation and as such do not benefit from PE.

This paper builds upon previous work by Gonçalves (2021) and presents an evaluation framework for source text informed by the features of Customer Support interactions, which is currently being put into production at Unbabel as a means of evaluating the quality of source language text. As well as the extent to which we are able to provide a methodology and a framework that can be used in the future to ensure accurate translation and improve the robustness of our MT models to source language noise.

In this paper we seek to address the following questions:

1. How can we adapt the prototype proposed in Gonçalves (2021) to better accommodate Customer Support translation in a production and business context?

2. Given the uniqueness of Customer Support content, how does the quality of the source affect translation quality?

2 Related Work

Noisy source text input is a common issue in MT and the translation quality of user-generated content has received some limited attention in literature. Vaibhav et al., (2019) for example, highlight the difficulties presented by source noise and introduce methods to improve MT system robustness to noisy source text from internet and social media, while Náplava et al., (2021) propose to statistically model errors from grammatical-error-correction with state-of-the-art NLP systems.

Regarding human evaluation of noisy source input, the research on error typologies and evaluation frameworks has mainly focused on the annotation of translation errors. There is a scarcity of investigation on the annotation of source text and limited work on the impact of source noise on MT output. As a consequence, research and guidance on the annotation of source text is equally lacking. Although there are no clear guidelines[1], we acknowledge the recent developments on the core MQM Framework (Lommel et al., 2014) to identify which categories of issue can be applied to both source and target errors. In addition, Tezcan et al. (2017) introduce the SCATE MT Error Taxonomy, which makes more of a clear distinction between the kinds of errors found in bilingual

[1] https://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html

and monolingual text. Notwithstanding the above, there is still very limited work on evaluation of source text independently of translation, particularly in the context of user-generated content and less so in the Customer Support use case.

Gonçalves (2021) considers the similarities and differences of the MQM Framework (Lommel et al., 2014), the SCATE MT Error Taxonomy (Tezcan et al., 2017) and proprietary error typologies developed internally at Unbabel. The same work introduces an adaptation of the MQM Framework (Lommel et al., 2014), and proposes a prototype of MQM-compliant typology for the annotation of source text errors, specifically tailored to user-generated content and Customer Support interactions. The proposed prototype included 4 parent categories and 28 terminal issues at 2 levels of granularity and was supported by specific annotation guidelines and decision trees to support annotation and the choice of the appropriate degree of severity of the selected errors. Still, the resulting typology showed a low agreement among annotators, meaning it lacked some robustness and wasn't efficient in a production setting.

3 Methodology

The primary goal of our revisions to the prototype typology presented in Gonçalves (2021) was to increase its robustness to Customer Support content and improve its effectiveness in a production setting. With this in mind, we made several adaptations which resulted in a new typology with 5 parent categories, 31 terminal issues and 2 levels of granularity. Where appropriate, we merged terminal nodes that resulted in low agreement and added a parent category. The full revisions to the initial prototype are detailed in the following section:

3.1 MQM Typology for Source Text Annotations

The adapted definitions from the typology prototype for the parent categories include the following:

Source Accuracy: While Accuracy is described as addressing the relationship between target and source text (Lommel et al., 2014), Source Accuracy addresses the mapping between A (actual source written by a customer or an agent on-the-fly) and B (intended source). This category is used when the semantic meaning or the conceptualization of an idea is compromised when, for instance, an agent or a customer does not finish a sentence and it is impossible to infer the intended meaning.

Fluency: In the MQM Framework (Lommel et al., 2014), Fluency includes issues related to the form or content of both source and target text. In this adaptation, Fluency addresses issues that affect the reading and comprehension of the text such as grammar, syntax and spelling. This category also determines how successfully the text can be interpreted as 'native-like' and that it would be understood to be such by a native speaker. Examples of this can be found in the wrong usage or the omission of prepositions, wrong function words or wrong choice of the appropriate verbal tense, mood or aspect.

Style: This category includes any stylistic issues found in source and target text (Lommel et al., 2014). In this adaptation, Style is to be used for stylistic issues, such as the use of register (e.g. formal vs informal), and specificities of online language, such as emoticons, conversational markers, idiomatic expressions or profanities.

Design and Markup: This category shares the same definition as the one provided in the MQM Framework (Lommel et al., 2014), although naturally with some differences in the issue types under it. It addresses any problem relating to design aspects (vs. linguistic aspects) of the content. One example of this is the segmentation of a complete sentence into several chat messages.

Locale Conventions: As defined in the MQM Framework (Lommel et al., 2014), this category is to be used when the text does not adhere to locale-specific mechanical conventions and violates requirements for the presentation of content in the target locale. This is related, for instance, to number, currency, addresses format used in a specific locale.

In order to facilitate the annotation process and aiming at high Inter-Annotator agreement and annotations consistency, as suggested by Artstein (2017), we provided annotators with an annotation decision tree[2] and a section concerning ambiguities.

[2] https://drive.google.com/file/d/1akddGjQbQHQEBxeBLFPSHwKsKGDT6MQ_/view?usp=sharing

3.2 Severities

A severity level indicates how grave or severe an error is. Having different levels of severity also helps to predict the impact of the source error text in the translation. We propose four different levels of severity: Critical, Major, Minor and Neutral (Lommel et al., 2014). It is also important to mention that, in order to facilitate annotation consistency, we also provided to the annotators a decision tree providing guidance on which severity is most likely to be suitable to the error or linguistic structure that is being annotated.

Critical: An error should be classified as critical when it contains information that may carry health, safety, legal or financial implications; a violation of geopolitical usage guidelines; a misrepresentation of the concerned company and their respective product/service; content that is completely inappropriate to its target audience and the meaning of the sentence is not understandable and cannot be inferred from the context.

Major: An error should be classified as major when there is misleading information; change of meaning and register wrongly used.

Minor: An error should be classified as minor when it impacts only minor aspects of meaning that can be resolved with proofreading.

Neutral: This label is not used for errors in the source text, but for linguistic structures that often have an impact on the quality or accuracy of the MT output. This includes only highly specific issue types: Emoticon, Segmentation, Conversational Marker, Idiomatic, Profanity, Abbreviation and Wrong Language Variety.

3.3 Annotation Rules

In addition to guidelines regarding the error definitions and the right degree of severity to be applied, we provided annotators with specific rules regarding the error span. We identified two main types of span, Continuous and Discontinuous, described below.

Continuous Span: This type of span involves a single continuous string of text. Based on their content, there are two sub-types of continuous span: single-word span and multi-word span. In single-word spans, a word is used incorrectly and only that item should be selected (e.g. misspelled word). On the other hand, in multi-word spans, an expression of more than one word in a continuous sequence is wrong. This usually applies to idioms or phrases that are assumed to be a single issue.

Discontinuous Span: These are errors involving a combination of two separate spans related to a single issue. Based on the relationship between the two spans, we can define four sub-types of discontinuous spans: delimiter spans, balanced spans, imbalanced spans, and asymmetrical spans. Delimiter spans are used to annotate typographic elements, balanced spans are used to highlight two disjoint but identical components of an issue when they are both incorrect, missing or added unnecessarily, imbalanced spans highlight two disjoint and distinct aspects of a single issue and, finally, asymmetrical spans are used to highlight an issue along with an element of context with which it is dissonant.

In order to support source text annotations for Customer Support, we included specific instructions to annotate the two sides of the Customer Support interaction, inbound (coming from the user) and outbound (from the agent) messages, also with specific instructions for user-generated content. These instructions were found to be important during the earlier annotations performed. The instructions were as follows:

1. Never annotate any Register issue type on inbound messages;

2. Do not annotate punctuation errors at the end of chat bubbles/messages;

3. Do not annotate capitalization errors in the beginning of a message.

Rule 1 is needed due to the fact that while agents are required to follow the register used by their company, users are not expected to do so, thus the Register issue type is irrelevant to inbound messages. Both rule 2 and 3 have exceptions: If the use of wrong punctuation at the end of a sentence changes its meaning, then it should be annotated; and if a capitalization error falls on a named entity, it should always be annotated with the Wrong Named Entity issue type.

Finally, in order to generate a production-ready typology to assess the quality of source text and by taking into account the improvements to the prototype mentioned above, we replicated the experiments in Gonçalves (2021) with the typology proposed in the latter work and measured the Inter-Annotator Agreement (IAA) calculated on a segment level, for both Customer Support and user-generated content data (as described in Section 4.1 below). To this end, we evaluated Cohen's Kappa Coefficient (Artstein, 2017) on the German corpus previously evaluated in Gonçalves (2021) and we observed an increased level of agreement across annotators, with an average Cohen's Kappa Coefficient of 0.5, versus the baseline showed in Gonçalves (2021) that exhibited an average Cohen's Kappa Coefficient of only 0.2.

4 Experimental Setup

The main application of source text annotation in a translation environment is to study and understand the propagation and the impact of source errors on the MT and PE steps.

In order to do so, we conducted three experiments with real client data in order to study how a particular communication medium influences agents and their communication and how MT systems handle user-generated content for chat message translation.

The first two experiments are client specific and, for privacy purposes, we refer to them hereafter as Client A and Client B respectively. The nature of the data from Client A is formulaic and repetitive. For this reason, we expect high quality translation results. The content chosen for this experiment was email threads, translated with a combination of MT and PE. We conducted a three-step alignment in the annotated data, where we made a comparison between the source text, the MT output and the post-edited target text of email threads. The language pairs analyzed were English to German ('en–de'), English to French ('en–fr') and English to Swedish ('en–sv').

On the other hand, Client B's data, being real-time chat, was translated with MT only. We included this particular client setting in order to study how chat communication affects the quality of the final MT output without any final human revision. The language pairs analyzed were English to German ('en–de') and English to Italian ('en–it').

Finally, in the third experiment, we randomly selected a sample of source texts coming from five different customers to study how user-generated content, in the context of chat conversations, impacts the final quality of the MT output. The language pairs analyzed in this experiment are Italian to English ('it–en') and Brazilian Portuguese to English ('pt–br–en').

4.1 Data and Data Preparation

In this section we present the data used for the experiments outlined in this paper, how they were translated and evaluated.

Corpus: Our main corpus is made up of 39,389 source text words across six language pairs, divided into three sub-corpora, each one corresponding to one specific experiment, as shown in Table 1.

Client A	
Language Pair	**Number of Words**
en–de	10,325
en–fr	14,520
en–sv	9,732
Client B	
Language Pair	**Number of Words**
en–de	1,288
en–it	1,261
User-generated	
Language Pair	**Number of Words**
it–en	1,088
pt–br–en	1,148

Table 1: Corpora sizes by number of words

Linguistic Resources: We applied customers' terminology to source texts, MT and PE translations and we provided our annotators and post-editors with specific customers style guides, language guidelines, the required formality level and also, in the case of annotators, the source text annotation guidelines produced in the context of these experiments.

Data Anonymization All data were anonymized in accordance with the European General Data Protection Regulation[3] (GDPR). Sensitive data and Personal Identifiable Information (PII) present in our corpus were identified using a proprietary Named Entity Recognition System (NER) and subsequently replaced with a placeholder tag.

MT Systems: The MT output analysed in the experiments presented in this paper was produced by different production MT systems pro-

[3] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&f20rom=EN

prietary to Unbabel. Unbabel's MT engines are transformer-based models (Vaswani et al., 2017) trained with the Marian toolkit (Junczys-Dowmunt et al., 2018). The models undergo varying levels of domain adaptation depending on several factors such as the client, language pair, and use case. The base models on which domain adaptation is applied are trained using millions of sentences of publicly available parallel data for a given language pair, from domains such as government and news. For our experiments, the user-generated content was translated by these "base" models, which underwent no domain adaptation. Email threads were translated with engines fine-tuned to tens to hundreds of thousands of parallel sentences of proprietary email content. Chat messages were translated with engines fine-tuned to tens to hundreds of thousands of parallel sentences of proprietary chat content specific to a single client.

Human Post-Edition: Unbabel's translation model is based on a combination of MT and human Post-Edition. In order to supply customers with a continuous customer support, Unbabel's post-editors are not necessarily professional post-editors, that would also entail a higher translation cost, but rather non-professional and bilingual. This allows Unbabel to grow and scale global communities and to provide human-corrected translations with very fast turnaround times.

Human Evaluation: Human evaluations were performed by Unbabel's PRO Community, made of professional translators and linguists with relevant experience in linguistic annotations and translation errors annotations. In order to properly assess translations quality, annotators must be native speakers of the target language and with a proven high proficiency of the source language, so that they can properly capture errors and their nuances. For the experiments outlined in this paper, the human evaluation was divided into two parts:

1. Source texts were evaluated with the adapted and improved Source Errors Typology outlined in this paper;

2. MT and PE outputs were evaluated by using the annotation framework adopted internally at Unbabel, which is an adaptation of the MQM Framework (Lommel et al., 2014) and that is tailored to assess Customer Support translated content.

5 Results

Our goal was to evaluate how errors present in source text impact the quality of the MT output and how they propagate and may be overlooked in human PE. This section presents the results obtained in the experiments outlined in Section 4. For simplification reasons, we refer to the experiment relating to Client A as "Client A Experiment", and similarly the experiment related to Client B, as "Client B Experiment" and, finally, the experiment run with user-generated content as the "User-generated content Experiment".

5.1 Client A Experiment

This experiment focused on the impact of repetitive content on the MT and PE targets. The MQM results for the source text and the two translation steps, as well the errors and their breakdown, are shown in Tables 2, 3, 4 and 5.

Language Pair	Source	MT	PE
en–de	79.15	94.8	90.2
en–fr	32.16	84.1	86.2
en–sv	-109.54	47.5	94

Table 2: Average MQM scores for Client A, Emails

Among the three language pairs, the most common error found is Code Switching, which refers to whenever another language, besides the source language, is used in the source text. This error was annotated as Critical because a source language native speaker, in this case English, would not understand messages written in another language. Qualitative analysis revealed that the language used in the source text was actually the target language. This is mainly due to the fact that agents did not have the right answers or templates to answer in English and they used pre-existing material available in the target language such as, for example, published FAQs and Knowledge Base articles. We observed that the MT systems are not robust enough to handle source text written in the target language, and, as this kind of template was used in a very repetitive way, this type of error occurred multiple times in the source data and was propagated to the MT outputs. The translation flow used to translate this content was MTPE and we observed that the poor MT output produced by Code Switching issues was correctly rendered by post-editors in the majority of cases.

Error	Neutral	Minor	Major	Critical
Code Switching	0	0	0	131
Punctuation	0	104	0	0
Capitalization	0	51	0	0
Omission	0	47	0	0
Segmentation	40	0	0	0

Table 3: Client A, top 5 errors with severity for en–de

Error	Neutral	Minor	Major	Critical
Code Switching	0	0	0	889
Segmentation	203	0	0	0
Punctuation	0	163	2	0
Whitespace	0	2	61	0
Omission	0	148	0	0

Table 4: Client A, top 5 errors with severity for en–fr

English–German: The Code Switching issues found in this language pair led to the MT engines generating errors in the target text. These were caused by additions and omissions of nouns and also the occurrence of a non-existing word. As a result, the information contained in the source text was altered in both MT and PE translations. Table 6, example (1) shows Code Switching errors in the source text that resulted in the substitution of a German word "Könnt" by a non-existing word, "Önnt", the change of the pronoun "ihr" ('you' in English) into the determiner "das" ('that' in English) which modified the meaning of the sentence, and the rephrasing of a sentence where there was an addition of a noun and a change of POS of a word that slightly altered its meaning, and an addition of the word "Rücksenders" which was unrelated to the rest of the sentence.

It is worth noting from table 2 that PE appeared to slightly degrade the MQM score. Whilst we generally conclude that PE will improve the translation quality, where the MT is already of a high quality, post edition can very rarely introduce noise.

English–French: Code Switching was, once more, the most common issue. Example (2) in Table 6 shows examples where this caused the addition of the word "numéro" ('number' in English) in the MT output. This example is very particular because the addition caused in the source text resulted in a better phrasing of the message conveyed.

Error	Neutral	Minor	Major	Critical
Code Switching	0	0	0	1,975
Segmentation	290	0	0	0
Punctuation	0	207	0	0
Capitalization	0	47	0	0
Omission	0	42	0	0

Table 5: Client A, top 5 errors with severity for en–sv

English–Swedish: This language pair had the highest occurrences of Code Switching annotations. In example (3) in Table 6, the source text was changed by the MT engine, affecting its original meaning. This created a semantic error in the target text, where the noun "ändring" ('change' in English) was changed to "service", an error that was not corrected in the PE translation.

5.2 Client B Experiment

In this experiment we aimed to study how the unique features of text generated in chat conversations outlined above, even in a more controlled environment such as Customer Support Centers, affect the quality of the MT output. The MQM results for the source text and the MT output, as well the errors and their severities, are shown in Tables 7, 8 and 9.

English–German: In this language pair, different errors occurred. In example (1) in Table 10, there was a Segmentation issue where the last letter ('e') of the noun "issue" was split into another chat message. This resulted in a critical error in the MT output, by leaving the segmented word "ISSU" untranslated and with the wrong capitalization. It is also important to note that this example shows how linguistic structures annotated as Neutral in the source text can produce critical errors in the MT output.

English–Italian: In example (2) in Table 10, the named entity "WhatsApp" was written in the source text with an extra whitespace and with the wrong capitalization ("whats app"). This resulted in a critical error in the Italian target text where the translation of this named entity was completely changed ("app quale"). With a whitespace separating this named entity, the MT translated both words separately and literally.

(1) **Code Switching (en–de)**	
Source	**Könnt** ihr mir einen retouren Aufkleber bitte schicken?
MT	**Önnt das** mir eine Retouren Aufkleber bitte schicken?
PE	**Önnt das** mir eine Retouren Aufkleber bitte schicken?
(2) **Code Switching (en–fr)**	
Source	Livraison manquante commande PHONENUMBER–0
MT	Livraison manquante de la commande **numéro** PHONENUMBER–0
PE	Livraison manquante de la commande **numéro** PHONENUMBER–0
(3) **Code Switching (en–sv)**	
Source	Re: Din **ändring** på PHONENUMBER–0
MT	Re: Din **service** på PHONENUMBER–0
PE	Re: Din **service** på PHONENUMBER–0

Table 6: Client A Experiment, examples of Code Switching

Language Pair	Source	MT
en–de	84	85.18
en–it	92.22	86.47

Table 7: Average MQM scores for Client B, Chat

Error	Neutral	Minor	Major	Critical
Punctuation	0	11	0	0
Omission	0	9	0	0
Capitalization	0	7	0	0
Segmentation	6	0	0	0
Word Order	0	0	5	0

Table 8: Client B, top 5 errors with severity for en–de

5.3 User-generated Content Experiment

In this experiment we focused on chat messages written by users to Customer Support agents not only to study the aspects of user-generated content in chat conversations, but also how they affect the MT output with no PE intervention. The MQM results for the source text and the MT output, as well as the errors and their severities, are shown in Tables 12, 13 and 14.

Brazilian Portuguese–English: Spelling errors, as expected, are among the most frequent issues annotated in chat messages written by users. Example (1) in Table 11 shows how the typo in the word "elçes" (which should actually have been "eles"), produces an untranslated critical error in the MT output.

Italian–English: As with the previous language pair, there were multiple minor errors and neutral linguistic structures that had an impact in the MT output through the propagation of major and critical examples. The idiomatic expressions are another mark of spontaneous speech used in chat messages and example (2) in Table 11 shows how idiomatic linguistic structures, annotated as Neutral, produce critical mistranslations in the MT output, where the idiomatic expression present in the source "mi sbatte fuori" (literally in English "it kicks me out"), was mistranslated into "it bangs me out".

Error	Neutral	Minor	Major	Critical
Punctuation	0	24	0	0
Capitalization	0	9	0	0
Wrong Named Entity	0	3	0	0
Addition	0	3	0	0
Omission	0	3	0	0

Table 9: Client B, top 5 errors with severity for en–it

Another mark of chat language is the usage of abbreviation. In the annotated data, abbreviations used in the source led to untranslated critical errors in the MT English target. Example (3) in Table 11 shows how the abbreviation "nn" of the Italian negation "non" produced an untranslated critical error in the English MT output.

Finally, it is worth mentioning an example of profanities found in the source data. Due to the fact that the customer support exchange can sometimes be stressful, users tend to express their frustration through the use of profanities. The profanity used

(1) Segmentation	
Source	I am sorry to hear about the **issu**
MT	Es tut mir leid, von der **ISSU** zu hören.
(2) Wrong Named Entity	
Source	When you share via any platform such as email or **whats app**.
MT	È Condividi tramite piattaforme come e-mail o **app quale**.

Table 10: Client B Experiment, English–German examples of Segmentation and Wrong Named Entity

(1) Spelling (pt–br–en)	
Source	mas consigo comprar com **elçes**?
MT	But can I buy with **elçes**?
(2) Idiomatic (it–en)	
Source	Oggi dopo l'aggiornamento inizia a caricare le partite e **mi sbatte fuori**..
MT	Today after the update starts loading the games and it **bangs me out**..
(3) Abbreviation (it–en)	
Source	Però io voglio capire perché **nn** riesco ad acquistare.
MT	But I want to understand why **nn** can buy.
(4) Profanity (it–en)	
Source	**Cazzo** ma parlo arabo?
MT	**Cazzo** but I speak Arabic?

Table 11: User-generated Content Experiment example errors

Language Pair	**Source**	**MT**
it–en	82.54	60.32
pt–br–en	91.11	27.03

Table 12: Average MQM scores for User-generated Content Experiment, Chat

Error	**Neutral**	**Minor**	**Major**	**Critical**
Spelling	0	17	7	0
Idiomatic	5	0	0	0
Wrong Named Entity	0	3	1	0
Omission	0	3	1	0
Segmentation	3	0	0	0

Table 13: User-generated Content, top 5 errors with severity for pt–br–en

in example (4) shown in Table 11 resulted in a critical untranslated error in the MT English target.

5.4 The Importance of Source Quality

Supplementary to the above analysis, we measured the Pearson's r correlation score at a document level between the MQM scores on the source text (measured using the typology presented in this paper) and the MQM of the resulting translations for Client A.

Error	**Neutral**	**Minor**	**Major**	**Critical**
Wrong Named Entity	0	23	0	0
Spelling	0	16	2	0
Whitespace	0	16	0	0
Punctuation	0	7	0	0
Idiomatic	7	0	0	0

Table 14: User-generated Content, top 5 errors with severity for it–en

Whilst we did not note a significant correlation for en-fr, we did however note Pearson scores of 0.38 and 0.33 for en-de and en-sv respectively (all significant to $p<0.05$).

From this we conclude that the effect of source noise on output translation quality is relatively pronounced. This further underlines the importance of source text quality in achieving high quality translation and the benefits of the framework presented in this paper as a means of measuring the same.

6 Conclusions

In this work we present an MQM-compliant annotation error typology that could be applied to evaluate the quality of source texts produced in a Customer Support environment that are translated with MT and PE. In particular we demonstrate the fundamental importance of source text quality to obtaining a high quality translation output. We further demonstrate how very specific source errors propagate to the MT targets, which generally lack robustness to these kinds of noise. It was also generally observed that in most MTPE translation flows, the PE step was beneficial to the final quality of the translation output with a resulting increase in the MQM scores. As machine translation is more widely deployed in a Customer Support context as a means of scaling service globally, the unique features of customer interaction with agents will continue to present unique challenges to MT. An obvious future direction for our work is in unifying approaches to improving MT (such as those in Vaibhav et al. (2019)) with our tailored framework as means of improving the robustness of MT to the benefit of the Customer Support use case. Equally, the same could be applied to mitigating the effects of human translation errors resulting from poor source quality.

Acknowledgements

This work was partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020 and through FCT and Agência Nacional de Inovação with the Project Multilingual AI Agent Assistants (MAIA), contract number 045909. Besides this, the authors want to thank João Graça, for believing in how source text quality is crucial for the MT domain, and to Christine Maroti and Amin Farajian, for their help and expertise on Machine Translation.

References

Artstein, Ron, 2017. *Handbook of Linguistic Annotation*, chapter Inter-annotator Agreement, pages 297–313. Springer.

Gonçalves, Madalena. 2021. Analysis on the impact of source text quality: Building a data-driven typology. *Repositório da Universidade de Lisboa.*

Höhn, Sviatlana, Alain Pfeiffer, and Eric Ras. 2016. Challenges of error annotation in native/non-native speaker chat. *Bochumer Linguistische Arbeitsberichte*, pages 114–124.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Lee, John and Stephanie Seneff. 2008. An analysis of grammatical errors in non-native speech in english. In *2008 IEEE spoken language technology workshop*, pages 89–92. IEEE.

Lind, Adam. 2012. Chat language: In the continuum of speech and writing.

Lommel, Arle, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463, 12.

Nasr, Alexis, Geraldine Damnati, Aleksandra Guerraz, and Frederic Bechet. 2016. Syntactic parsing of chat language in contact center conversation corpus. In *Annual SIGdial Meeting on Discourse and Dialogue*, pages 175–184.

Náplava, Jakub, Martin Popel, Milan Straka, and Jana Straková. 2021. Understanding model robustness to user-generated noisy texts. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 340–350. Association for Computational Linguistics.

Roturier, Johann and Anthony Bensadoun. 2011. Evaluation of mt systems to translate user generated content. In *Proceedings of Machine Translation Summit XIII: Papers.*

Rozovskaya, Alla and Dan Roth. 2010. Annotating esl errors: Challenges and rewards. In *Proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications*, pages 28–36.

Sinha, Avanika, Niroj Banerjee, Ambalika Sinha, and Rajesh Kumar Shastri. 2009. Interference of first language in the acquisition of second language. *International Journal of Psychology and Counselling*, 1(7):117–122.

Tezcan, Arda, Véronique Hoste, and Lieve Macken. 2017. Scate taxonomy and corpus of machine translation errors. *Trends in E-tools and resources for translators and interpreters*, pages 219–244.

Vaibhav, Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. *arXiv preprint arXiv:1902.09508.*

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Guyon, I., U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.