Data-driven Asian Adapted MQM Typology and Automation in Translation Quality Workflows

Beatriz Silva, Unbabel
Marianna Buchicchio, Unbabel
Daan van Stigt, Unbabel
Craig Stewart, Phrase
Helena Moniz, University of Lisbon; INESC-ID; Unbabel
Alon Lavie, Phrase

ABSTRACT

In this study we aim to test the impact of applying translation error taxonomies oriented towards European Languages in the annotation of Asian Languages. We aim to demonstrate how an error typology adapted for the latter languages can not only result in more linguistically accurate annotations, but also how this can be applied to automating and scaling translation quality evaluation.

As such, we propose a Translation Errors Typology that aims to cover the shortcomings of the Multidimensional Quality Metrics (Lommel et al. 2014) framework (MQM) in what concerns the annotation of the East Asian Languages of Mandarin, Japanese and Korean. The effectiveness of the typology here proposed was tested by analysing the Inter-annotator agreement (IAA) scores obtained, in contrast with the typology proposed by Ye and Toral (2020) and the Unbabel Error Typology[1]. Finally, we propose a way of automating Translation Quality Workflows through a Quality Estimation (QE) technology that is able to predict the MQM scores of machine translation outputs at scale with a fair correlation with the human judgement produced by applying the East Asian Languages MQM module proposed in this study.

KEYWORDS

Inter-annotator agreement, multidimensional quality metrics, translation quality workflows, annotation procedures of quality processes, East Asian languages, quality estimation.

1. Introduction

Manual annotation of translation errors has been used in recent years simultaneously with automatic metrics such as BLEU (Papineni *et al.* 2002) and COMET (Rei *et al.* 2020) as a manual quality assessment process for MT outputs. While automatic quality metrics allow fast evaluations and demand low human resources, it is difficult for them to accurately imitate the behaviour of human annotators. On the other hand, manual quality assessment processes require human effort, which makes them slower, more expensive and often inconsistent as annotators can be in disagreement with each other and even with themselves throughout different annotation jobs (Macháček and Bojar 2015: 85). However, unlike most automatic metrics, they have the advantage of not being dependent on reference translations (Lommel *et al.* 2014a: 165). Simultaneously, they also constitute a more detailed form of error analysis which allows the identification of the exact problems existing within a translation which results in data that is essential for the improvement of automatic metrics (Popović and Arčan 2016: 27). Despite automatic metrics being trained to have close correlation with human judgements, they still lack aspects that human assessment processes can provide, such as the identification of span

https://doi.org/10.26034/cm.jostrans.2024.4713

and classification of error types, which automatic metrics alone do not allow, making them insufficient in the analysis of translation errors.

The Multidimensional Quality Metrics (MQM) framework, developed by Lommel *et al.* (2014) in the context of the QTLaunchPad project[2], is currently considered to be the standard for manual quality evaluation in the translation industry. Due to the fact that the MQM framework was built to be customizable, it allows Language Service Providers (LSPs) and other entities who intend to use it to create typologies fit to their specific needs. As recommended by the MQM framework developers, MQM-derived typologies should not be overly fine-grained so as to not overload annotators with information (Lommel 2018a). At the same time, typologies that are too general cannot provide a clear image of what is actually wrong with a translation and, thus, what needs to be improved in MT engines (Vilar *et al.* 2006: 697). In this context, it is challenging to create a typology that is equally efficient for annotating very distinct languages. This results in the fact that LSPs that work with many different languages work with error annotation typologies that suit some languages better than others. In this article we focus on observing the consequences of using typologies conceived from a general point of view to annotate certain sets of languages, specifically in relation to translation from English to the East Asian Languages of Japanese, Korean and the simplified and traditional variants of Mandarin Chinese[3]. In Section 3 of this article, we explore the methodology for building a typology for annotation of translation errors adapted for the languages aforementioned based on the issues existing within annotations performed with a European Languages-oriented Typology, which does not contain some fundamental issue types for annotation of those languages, as well as appropriate guidelines to address specific situations and problems that arise when annotating language pairs (LPs) with them as the target language. For example, in the case of languages such as Japanese or Mandarin Chinese which do not use whitespaces, it is important to have guidelines to guide the annotators on how to handle errors such as *Omission* in order to create a rule that allows consistent annotations. Moreover, having issue types adapted for the translation direction being annotated can result in more accurate evaluations of the problems in a translation.

We aim to prove that this typology can address the shortcomings MQM presents in the context of annotation for East Asian Languages, namely the lack of fine-grained issue types to accurately reflect translation problems concerning these languages, as well as the lack of detailed guidelines for the usage of issue types which may be ambiguous during the annotation process, such as the case of *Omission* errors mentioned above. We also intend to demonstrate that in addition to being used for annotation of translation errors, this typology can simultaneously be applied in the context of automation of quality processes for MT in order to address another shortcoming of MQM and overall manual quality assessment processes, which is the fact that they are costly and time-consuming. The importance of having a typology that is adequate for the annotation of East

Asian Languages in the context of automation of annotations is that our MQM-QE[4] models are trained to imitate human evaluation behaviour and, as such, it is essential that the data used to train them is as consistent and accurate as possible, which is difficult to obtain from data from typologies that do not suit the languages being annotated. As such, in addition to the new issue types featured in the proposed typology, we developed specific annotation guidelines with sections dedicated to tricky cases and decision trees, both for the selection of the correct issue types and their corresponding severity, with the purpose of reducing ambiguity and increasing Inter-annotator agreement (IAA) scores, which are all important for training our MQM-QE model.

2. Manual Quality Metrics and Typologies for Annotation of Translation Errors

Manual quality metrics are essential to allow the identification of translation errors at a level that automatic metrics cannot guarantee. The Direct Assessment (DA) approach (Graham *et al.* 2013), as explained by Ma *et al.* (2017: 599) is a method that relies on crowd-sourcing methods to measure how much a translation hypothesis relates to a reference or the source text on a scale from 1 to 100. This approach has the disadvantage of not being performed by considering a standardised set of error types and, as such, the data obtained from it is not fine-grained and cannot reflect the exact errors an MT output presents. The translation rating approach, on the other hand, is a method through which translations are rated according to a Likert scale and usually in terms of *Fluency* and *Accuracy* (Snover *et al.* 2009: 259). However, as pointed out by Koehn and Monz (2006: 110), this metric is difficult to apply, due to the fact that human evaluators struggle to attribute these scores to the quality of a translation. In this work we focus on MQM, which as mentioned before is the current standard for fine-grained manual translation quality evaluation and, specifically, on proposing an MQM-compliant typology that is adapted for East Asian languages.

The main objective of creating the East Asian Languages MQM module we propose in this article is to have a Manual Translation Quality Assessment (TQA) process which can be used reliably in the evaluation of MT quality for these languages. The appearance of the first systems of lists of translation errors with severities started between the 1990s and the 2000s with the creation of the LISA QA Model and the SAE J2450. The development of the LISA QA Model, in particular, was fundamental for the later development of the MQM framework, which is considered the current standard for manual quality evaluation in a wide range of translation industries, as it featured a list of possible translation errors and a total of three attributable severities in a structure similar to MQM. After the dissolution of LISA, in 2012 the Translation Automation User Society (TAUS) created the DQF (Dynamic Quality Framework), which contained a total of six error types which were further divided into subtypes and four severity levels, in a structure similar to the MQM framework (Lommel

2018b: 123-124), created in 2015 for the QTLaunchPad research project. As mentioned above, the MQM framework was based on the structure of the LISA QA Model but it also aimed to solve the problems its one-size-fits-all approach presented, by allowing MQM to be customizable (Lommel *et al*. 2014b: 459-460). More recently, a more simplified version of MQM, MQM Core[5], was also developed and in 2022 a new version of MQM[6] was released.

Prior to discussing the MQM-compliant typology we propose, it is relevant to analyse other customised typologies derived from MQM, specifically typologies applied in the annotation of the East Asian Languages our study covers.

One of these typologies is the Unbabel Error Typology[7], which is used in a business context for the annotation of several LPs, including East Asian Languages. As per the recommendations of MQM creators, it is a typology that attempts to avoid being overly fine-grained. However this means that it favours certain sets of languages more than others, namely European languages, neglecting issue types that are essential when annotating East Asian Languages such as issue types for particles and classifiers that would be too specific to include in a general typology.

In the context of annotation typologies applied to East Asian Languages, the typology proposed by Ye and Toral (2020) is of great importance, as it was created with the goal of being used for annotation of the English to Mandarin Chinese LP and, as such, contains specific issue types that are relevant for this translation direction, including issue types for errors with classifiers and particles.

In order to test the effectiveness of the typology we propose in this article and its potential to solve the problems that we verified when analysing data annotated with a general typology, we conducted a testing study in which we annotated the same content for each LP with the two typologies mentioned above, in addition to the one we created. The comparison with these two typologies is especially relevant due to the fact that even though both are MQM-compliant taxonomies, they are opposites in the sense that one, the Unbabel Error Typology, is a general taxonomy conceived with the purpose of being used for the annotation of several LPs; while the other, the MQM-compliant typology proposed in 2020 by Yuying Ye and Antonio Toral, is a typology that was created specifically for annotation of translations in the English to Mandarin Chinese direction. It should be noted that although this typology was proposed only to be used for annotation of the aforementioned LP, the issue types concerning particles and classifiers are very relevant in the annotation of Korean and Japanese as well, as noted by the annotators for these languages.

Aside from the issue concerning the annotation of particles and classifiers, the most significant difference between these two typologies is their structural organisation particularly in the annotation of *Mistranslation* and *Function Words* errors, as represented in Tables 1 and 2, which also proved to be important in the testing phase that will be discussed in Section 3.2., since it impacted the IAA scores that were used to measure and compare the performance of each of the three typologies. It is necessary, then, to briefly discuss these typologies with special focus on the aspects that make them either inadequate or appropriate, respectively, for annotating the LPs under discussion in this article.

Mistranslation		
Unbabel Error Typology	East Asian Languages MQM Module	Ye and Toral (2020)
Named Entity	Wrong Named Entity	Entity
Overly Literal	Overly Literal	Overly Literal
Lexical Selection	Lexical Selection	
Ambiguous Translation	Transliteration	
False Friend		
Shouldn’t Have Been Translated*****		
Spelling*****		
Wrong Date/Time		
Wrong Number		
Wrong Unit Conversion		

Table 1. Organization of *Mistranslation* issue types. The shaded cells represent the error categories that the typologies have in common.

Function Words			
Unbabel Error Typology	East Asian Languages MQM Module	Ye and Toral (2020)	
Fluency	Linguistic Conventions	Fluency	
Grammar → Function Words			
Wrong Preposition	Wrong Preposition	Extraneous	Preposition
Wrong Conjunction	Wrong Conjunction		Adverb
Wrong Pronoun	Wrong Pronoun		Particle
Wrong Auxiliary Verb	Wrong Auxiliary Verb	Incorrect	
Wrong Determiner	Wrong Particle	Missing	
	Wrong Classifier		

Table 2. Organization of *Function Word* issue types. The shaded cells represent the error categories that the typologies have in common.

Our typology was designed to only allow the selection of the end nodes for each error and to block the selection of all parent nodes. On one hand, this structure can have a positive influence on IAA scores, as it results in a lower number of total selectable issue types and reduces ambiguity, and, on the other hand, results in annotations that contain more detailed information about the translation errors in a text, in contrast with what would be the case if the selection of a broad category such as *Mistranslation* was allowed. Even though the error typology created by Ye and Toral (2020) contemplates the selection of *Mistranslation* although it is not an end node, in order to have a consistent structure during our annotation experiments, we disabled the selection of the parent node of *Mistranslation* for all typologies.

It should also be mentioned that the error typology we propose was built from within Unbabel and, as such, its structure is similar to that of the Unbabel Error Typology, as can be noted in Tables 1 and 2. This means that these two typologies share an identical structure and, specifically in relation to the annotation of function words, are organised in a scheme which separates grammatical errors that occur upon them from their omission, which for both typologies is included under the mother category of *Omission*. Contrary to this scheme, the typology proposed by Ye and Toral (2020) contemplates both the wrong usage and the omission of function words as grammatical errors and, as such, both types of errors are included under the *Function Words* (*Grammar*) category. This was an important factor to consider when evaluating the IAA scores obtained with the typology we propose, as the annotators that participated in the experiments were already familiar with the Unbabel Error Typology as well. On the other hand, however, it is also relevant that categories such as *Shouldn't Have Been Translated* and *Spelling* (*), included under *Mistranslation* in the Unbabel Error Typology, are also included in the East Asian Languages MQM module but under different coarse categories. Similarly, the issue types of *Wrong Date/Time*, *Wrong Number* and *Wrong Unit Conversion* were all included under *Wrong Named Entity* in the East Asian Languages MQM module.

3. Building an error typology

In order to build the error typology adapted for East Asian Languages, we based our decisions on a data-driven approach. As such, annotation data from these languages previously annotated with a general error typology was analysed[8]. The aim of our study in this phase was to identify the errors produced during the annotation process in the four East Asian Languages analysed in this study: English-Korean, English-Japanese, English-Traditional Chinese and English-Simplified Chinese. During this phase, we produced a detailed analysis of the errors and determined which among these could be attributed to the fact that the typology used for annotation was not entirely adequate for the languages at hand. This would provide a basis for understanding what should and should not be included in a typology adapted for East Asian Languages, as well as what type of specific guidelines would be needed in an effort to reduce annotation subjectivity.

In this phase we evaluated 600 to 1200 annotated segments per LP while considering three types of annotation errors which constitute the parameters where most disagreement between annotators can be found (Lommel *et al.* 2014c: 2). Such errors are considered as follows in our analysis:

- Typology Errors: if the translation error was annotated with the wrong issue type or if the error should not have been annotated;
- Span Errors: if the error was annotated in the wrong place or spanning an incorrect length;

- Severity Errors: if the wrong severity was attributed to the error[9].

For this experiment, all the segments in the dataset contained errors which were annotated at document level on machine translated emails in the Customer Service domain. The annotations were performed by experienced annotators[10] and evaluated by an expert linguist in terms of the three parameters described above. Upon this evaluation, we found that there were annotation errors across all categories, as indicated in Table 3.

	Typology Errors	Span Errors	Severity Errors
All	**35.6%**	13.2%	23.8%
Japanese	**20.4%**	19.5%	18.4%
Korean	**59.4%**	14.0%	11.6%
Simplified Chinese	15.8%	9.6%	**19.9%**
Traditional Chinese	7.2%	1.2%	**43.9%**

Table 3. Percentage of types of annotation errors with the Unbabel Error Typology. The numbers in bold highlight the most common type of annotation error per LP.

As seen in Table 3, the most visible type of mistakes are "Typology Errors", with 35.6% of all the errors annotated across all LPs containing this type of issue This means that the annotators were frequently using the wrong issue type for identifying certain errors and that they were also annotating errors unnecessarily. A more detailed analysis of the types of typology errors in this dataset revealed that many were related to the misannotation of *particle* and *classifier* errors. For example, in the case of Japanese and Korean, particle errors were frequently annotated as preposition errors. This is problematic not only due to the fact that it is agrammatical and, thus, reduces the value of the information obtained from these annotations, but also because the non-existence of an appropriate issue type for annotating these errors meant that their annotation was not consistent and that other issue types were also attributed in addition to preposition issue types, affecting the uniformity of the annotations and, consequently, their reliability. The same is true in the case of classifiers, which were annotated as determiner or other part-of-speech (POS) errors, mainly in the Simplified Chinese dataset.

While these were not the only annotation mistakes related to the selection of the wrong issue types, together with cases of transliteration they were the types of errors we believed were possible to avoid with an adapted typology. *Transliteration* was also considered as a relevant category to be added to our East Asian Languages MQM module due to the fact that such errors were annotated using different issue types, such as *Overly Literal* and *Lexical Selection*, which resulted in inconsistencies.

3.1. The East Asian Languages MQM module

In addition to the inclusion of issue types for annotation of the errors mentioned in the section above in the East Asian Languages MQM module, it was believed that the inconsistency in the annotation of other issue types which were frequently annotated inconsistently inter- and intra-annotator wise could be improved through specific guidelines which attempt to reduce subjectivity during the annotation process. In this section we present the annotation guidelines for the module as well as the module itself, followed by the results of the annotation experiments performed with this typology.

3.1.1. Annotation Guidelines

In order to build the annotation guidelines, we conducted several interviews with professional linguists working as translation errors annotators. We interviewed them on the main challenges and difficulties of choosing either the correct error span, issue type or severity. After gathering this feedback, we produced the annotation guidelines proposed in this study and ran an annotation pilot with test data. During the test, annotators were asked to provide additional feedback and we maintained contact with them to clarify doubts that arose during this process. After this first pilot phase, we were able to produce the final version of the guidelines proposed in this study. In other words, these guidelines were built upon several feedback cycles with the annotators and the pilot experiments on annotating data. As a result, the guidelines built for this typology are focused on providing the annotators sufficient instructions with the purpose of making the annotation process as objective as possible, not only in terms of disambiguation of issue types but also in relation to the selection of the correct span for each error and providing clear guidance on how to choose appropriate severities.

It should be noted that in the context of enabling automatic annotations for these languages as a final goal, which requires the proper identification of errors, it is essential not only to have the correct issue types associated with each error but also the correct span, as the proper identification of the error depends on it. At the same time it is also of significant importance to have the correct severities attributed to the errors, as they have a great impact on the MQM scores. Accordingly, in the guidelines for our typology adapted to East Asian Languages, we include detailed decision trees[11] for the selection of each issue type and the correct severity. In addition, along with definitions and examples for all issue types, the guidelines also include complementary information in order to avoid ambiguities. For example, the definition for *Wrong Named Entity*, as seen in Table 4, contains additional information on the cases in which a named entity should or should not be annotated.

Issue Type	Definition
Wrong Named Entity	The target contains an error related to named entities (names, places, etc.). The named entity may not match between source and target or it contains an error in the target, such as capitalization or orthography. This issue should **also** be applied when the named entity has been transliterated, unless the transliteration was required. **Please note** addresses and numerals should also be considered named entities and should be annotated as such in case they are wrong in anything other than their format.

Table 4. Definition for the *Wrong Named Entity* issue type in the guidelines for the East Asian Languages MQM module

The guidelines also contain a Tricky Cases section, where ambiguous annotation cases are addressed with instructions on how to proceed with them and illustrative examples, especially in terms of span selection, as seen in the example in Table 5.

Tricky Case	Instructions
Annotating Omission errors when there's no whitespace	When an Omission error occurs and there is no whitespace to annotate it on, the error should be annotated on the unit immediately after where it would be. In the case of a punctuation mark that has been omitted, the reasoning used for annotation is the same, only it should be tagged as Punctuation. However, if the omitted unit is bound to a specific word in the segment, for example in the case of particles, the error should be annotated on that word. **Note:** the error should be annotated on a full unit and not just on its first/last character.

Table 5. Tricky Case example from the guidelines for the East Asian Languages MQM module

More specifically, in addition to providing disambiguation instructions in relation to span selection issues, as in the example in Table 5, it attempts to clearly establish the difference between certain issue types, such as the distinction between *Lexical Selection*, *Overly Literal* and *Transliteration*, the latter of which is one of the new issue types the typology introduces. Similarly, it contains instructions on how to annotate with the new issue types for particles, which were believed to be likely to cause ambiguity due to the fact that they are attached to other words. Finally, based on difficulties previously observed in annotation data obtained with the

Unbabel Error Typology, a detailed explanation on how to annotate verbs was also included with examples and specific details for each of the languages.

3.1.2. New issue types

The full typology adapted for East Asian Languages, represented in Figure 1, contains a total of 39 selectable issue types, divided into 7 coarse categories, 24 daughter issue types, 13 granddaughter issue types and 6 grand granddaughter issue types.

The error typology adapted for East Asian Languages that we propose attempts to find a balance between being too fine-grained and overly simplified. As mentioned before, error typologies that are overly simplified provide little information about the exact errors that exist within a translation. As such, our error typology expands on the categories which frequently require additional information, such as *Mistranslation*, *Omission* and *Grammar*, which are also the categories the five new issue types we propose were added into. The new issue types, which are highlighted in Figure 1, are defined as follows in the annotation guidelines for the typology:

- Omitted Particle: a particle is missing in the target text;
- Omitted Classifier: a classifier is missing in the target text;
- Wrong Particle: a particle is used incorrectly (another particle should have been used instead);
- Wrong Classifier: a classifier is used incorrectly (another classifier should have been used instead);
- Transliteration: a term in the target has been transliterated instead of being accurately translated.

We expected these issue types would be transversal to all four LPs since all contained annotation mistakes related to these categories in previous annotation data and that, as such, it is possible to have one typology that is equally applicable to the four LPs. Similarly, we avoided the inclusion of issue types that would not be used in the annotation of these languages, such as *Agreement*, *Capitalization* and *Diacritics*. This reduces the information the annotators have to assimilate and, in parallel, serves the purpose of avoiding mistakes. For example, it is expected that the only situations where words remain in latin script in the target are when the text contains named entities, such as company names. In these situations, the named entity may be using the wrong capitalization in the target, in which case the *Wrong Named Entity* issue type should still be applied, as per the guidelines. In addition to these instructions, which may be overlooked by the annotators in some situations, an annotation error can be avoided by not including the *Capitalization* issue type in the typology

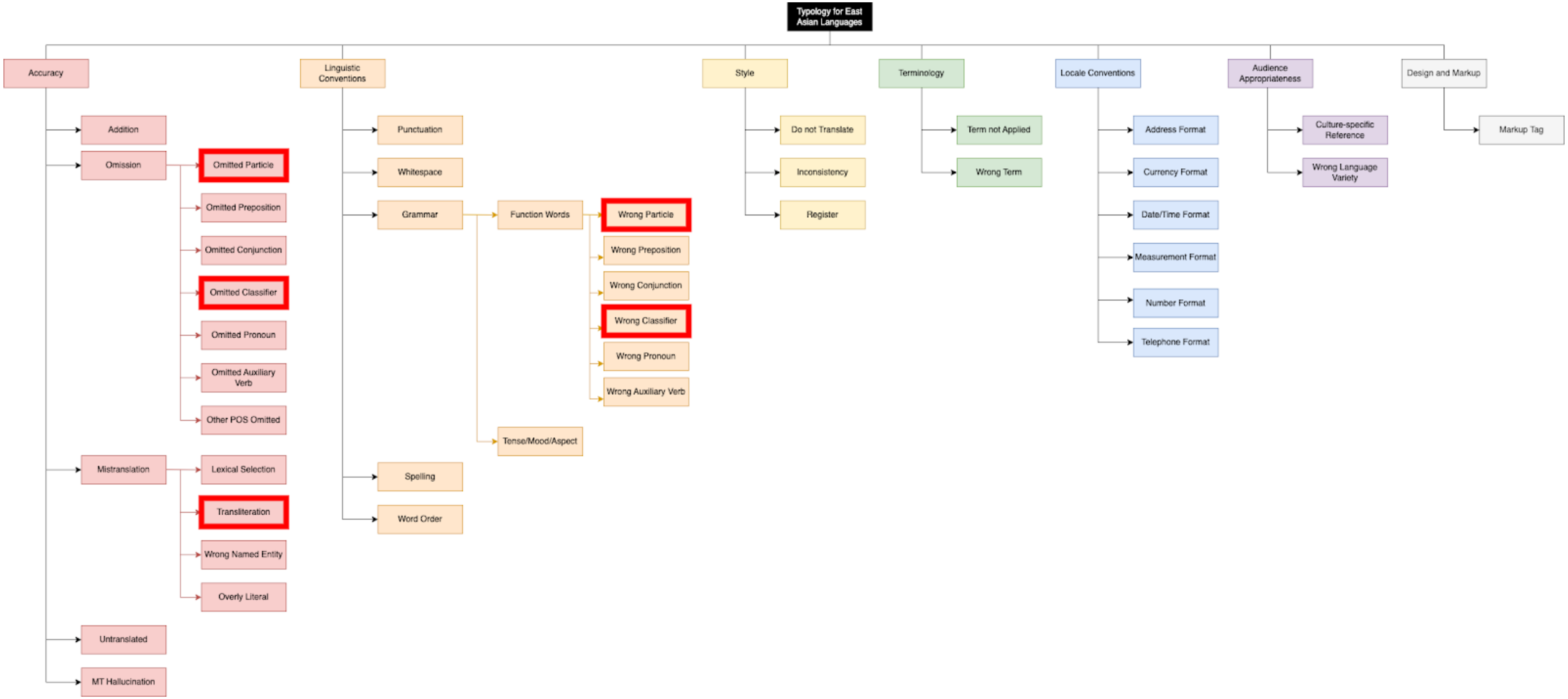


Figure 1. The East Asian Languages MQM module. The red boxes highlight the categories exclusive to the East Asian Languages MQM module

3.1.3. Results

As mentioned in Section 2, in order to test the effectiveness of our proposed typology we performed an annotation comparison study through which we could compare its performance against two other typologies in terms of Inter-annotator agreement (IAA), which can illustrate the clarity of a typology in terms of its issue types and guidelines (Lüdeling and Hirschmann 2015: 148-149). The annotation setup for this experiment consisted of two annotators per LP, one native speaker and one non-native speaker, annotating the same content three times, once with each typology being tested, with approximately a month of interval between each phase. Since in the business use case our annotation module applies to there are often non-native speakers working on annotation tasks and the interpretation of the typology may be different between native and non-native speakers, it was important to have annotations from both perspectives. For each LP the number of annotated segments depended on the data that was available, causing some disparity in the number of annotated words for each LP, the lowest number being for Japanese with approximately 1100 words and the largest corresponding to Simplified Chinese, with approximately 4900 annotated words. The annotations were performed on machine translated chat content in the Customer Service domain at document level. The first batch of annotations was conducted using the Unbabel Error Typology, since it required no further training from the annotators. The second batch was annotated with the typology proposed by Ye and Toral (2020), as it was the most different in terms of issue types and structure in order to distance the annotators from the Unbabel Error Typology structure before annotating the third and final batch with the East Asian languages annotation module, which resembles the former. The total of errors annotated during this experiment were distributed as represented on Table 6.

LP	**Unbabel Error Typology**		**Ye and Toral (2020)**		**East Asian Languages MQM module**	
	Annotator A	Annotator B	Annotator A	Annotator B	Annotator A	Annotator B
EN_JA	37	56	30	38	37	51
EN_KO	736	258	223	23	713	159
EN_ZH-CN	318	164	410	224	428	404
EN_ZH-TW	158	62	118	138	154	168

Table 6. Distribution of errors per typology and annotator

		Unbabel Error Typology	**Ye and Toral (2020)**	**East Asian Languages MQM Module**
Avg. batch IAA	EN_JA	0.628	0.526	0.526
	EN_KO	0.366	0.454	0.355
	EN_ZH-CN	0.464	0.352	0.520
	EN_ZH-TW	0.192	0.194	0.189
% of translations above 0.4 Cohen's κ	EN_JA	54.5%	58.3%	66.7%
	EN_KO	50%	8.3%	31.3%
	EN_ZH-CN	60%	28.6%	68.8%
	EN_ZH-TW	17.6%	5%	5%
Total number of issue types		47	15	39

Table 7. IAA scores and number of issue types per typology[12]

Table 7 illustrates the IAA scores and the number of issue types for each typology. As per Amidei *et al.* (2019: 347), according to previous investigations 0.40 Cohen's κ is the average IAA score obtained from human annotation, which is considered as fair to moderate rating defended by Landis and Koch (1977) (as cited in Amidei *et al.* 2019: 347). As such, we established a threshold where IAA scores of 0.40 Cohen's κ or above were considered as satisfactory. In addition to the overall IAA scores per LP, we also evaluated whether the percentage of jobs above the satisfactory threshold changed according to the typology being used.

In light of the results obtained, it can be affirmed that the typology we propose had positive results, especially in the case of Traditional Chinese, where it obtained the highest overall IAA score of all three typologies for this LP, and Japanese, where it had the highest percentage of jobs above the acceptable threshold. However, the IAA scores for Korean fell below the desired results for the typology adapted for East Asian Languages, while the same occurred with Simplified Chinese across all typologies. In the case of Korean, it is important to note that the difference in number of annotations between annotator A and B is substantial. This is due to the fact that the Korean data was heavily polluted with whitespace errors, which were not equally acknowledged by both annotators, and formatting errors which led annotator B to discard jobs from annotations rather than annotating them. When comparing the results closely, it is evident that there is not one typology that distinguishes itself with significantly overall higher scores in relation to the other typologies. It was important, then, to analyse the annotations in detail in order to understand the IAA scores and the advantages and limitations of each typology, in an effort to pinpoint what still needs improvement in our typology and how to approach this process.

The Unbabel Error Typology had the advantage of being familiar to the annotators. However, as seen in the last row of Table 7 it is the most extensive of the three typologies and it notably misses specific issue types for the annotation of the LPs under discussion.

In the case of the typology proposed by Ye and Toral (2020), it is a typology that was conceived for the annotation of English to Mandarin Chinese translations, which is one of the LPs our proposed typology aims to cover. As such, it contains issue types relevant for the annotation of said LP, which we concluded are transversal to the other LPs under discussion in this article. In fact, as seen in Table 7, its highest IAA scores correspond to the annotation of Korean and Japanese. In addition, it is the most reduced of the typologies, containing less than half the number of the issue types existent in the other two typologies. However, it lacks issue types for *Whitespace* and *Register*, which are essential in a typology to be applied to these four LPs. Additionally, the structure of the typology originated ambiguity issues, especially due to the fact that there is one other issue type apart from *Omission* to annotate missing function words (*Missing*), as seen in Table 2. Finally, this typology also performed poorly in terms of annotation of *Mistranslation* errors. In order to obtain detailed information about translation errors, we did not allow the selection of parent nodes during this annotation process. However, in the case of Ye and Toral's typology, this revealed to be a problem due to the fact that as per Table 1, the *Mistranslation* category in Ye and Toral's typology only contains two sub-categories, which are not applicable to all cases of mistranslation, forcing the annotators to choose other issue types such as *Unintelligible*, often not in agreement with each other.

Finally, our proposed typology had the advantage of containing specific issue types and guidelines for the annotation of the LPs under discussion. A detailed analysis of the annotations revealed that although the IAA results obtained with our typology were not as high as desired, aside from the issue type for *Transliteration*, which caused some confusion and, thus, disagreement between the annotators, the new issue types in the typology were successfully applied. This was true both in terms of unifying annotations, as seen in example (1), and transitioning the annotation of certain errors, such as particles, to the correct issue types, as seen in examples (2a) and (2b):

(1) (EN) Usually you can only see 1 wifi name on it.
(ZH-CN) 通常您只能在上面看到1**[Ø]无线网络**名称。

→ Unbabel Error Typology: Omitted Determiner
(Annotator A) / Other POS Omitted (Annotator B)
→ Ye and Toral's Typology: Missing Function Word
(Annotator A) / Classifier (Annotator B)
→ East Asian Languages MQM module: Omitted Classifier
(Annotator A and Annotator B)

(2)
(a) (EN) We have successfully cancelled the recurring payment with [PRODUCT].
(JA) [PRODUCT]**で**定期支払いをキャンセルしました。

(b) (EN) What is the [PRODUCT] ALPHANUMERICID-0 Dual-Band Smart Wi-Fi Wireless Router?
(KO) [PRODUCT] ALPHANUMERICID-0 듀얼 밴드 Smart Wi-Fi 무선 공유기**은** 무엇입니까?

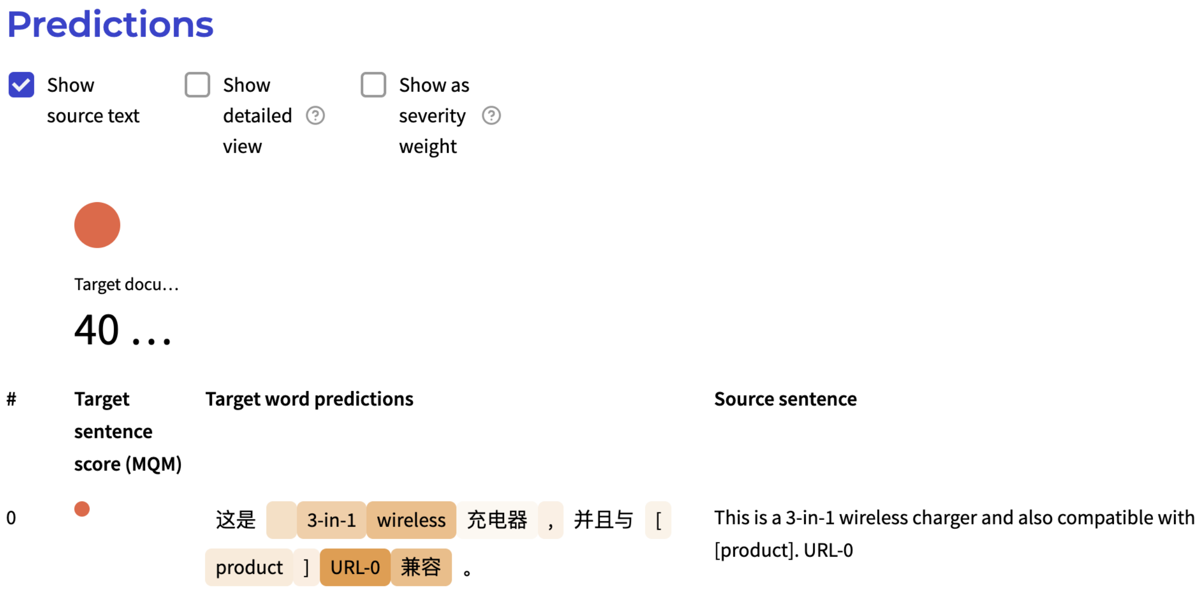
→ Unbabel Error Typology: Wrong Preposition
→ East Asian Languages MQM module: Wrong Particle

On the other hand, we found that its biggest limitation was the category of *Mistranslation*, which is divided into four subcategories which the annotators disagreed on how to use.

It is necessary to mention at this point that IAA scores and consistent annotations are fundamental in the context of the automation of annotations which will be discussed in the following section, as this process depends on already existing data. As such, it is important that this data is as accurate as possible.

4. Automation of Quality Workflows: MQM-QE for East Asian Languages

One of the main shortcomings of human evaluation methodologies and, in particular, the MQM framework, is that the evaluation performed by human annotators is slow, costly and particularly time-consuming. As a consequence it is now fairly common to seek automated methods of evaluation as a means of assessing translation quality at scale. In order to overcome this problem, and as Kepler *et al*. (2019:117) state, Quality Estimation (QE) provides the missing link between the human and the machine. We have built a suite of AI tooling for this purpose; in particular we leverage QE, an automated means of evaluating translation quality by presenting a source text and its corresponding translation to the MQM-QE model[13] here proposed and having it generate an MQM-like score at the end. Even though MQM-QE is an autonomous module, the predictions it generates are important to validate the results obtained in our research. Specifically we use a proprietary neural network which predicts the span and severity of the error in a given text by labelling each of the target tokens with either “OK” or “BAD” and automatically produces an MQM score based on the predictions. This allows us to identify not only the general quality of the text, but also the source of the errors responsible for degradation of the output score.



Predictions

☑ Show source text □ Show detailed view □ Show as severity weight

●

Target docu…

40 …

#	Target sentence score (MQM)	Target word predictions	Source sentence
0	●	这是 3-in-1 wireless 充电器 ， 并且与 [product] URL-0 兼容 。	This is a 3-in-1 wireless charger and also compatible with [product]. URL-0

Figure 2: MQM-QE User Interface

Figure 2 shows an example of an English to Simplified Chinese machine translation output and the interface used to analyse the quality predictions of the MQM-QE module proposed in this study. For illustration purposes we show our MQM-QE interface that is used to visualise the general quality of a translation, as well as the location of the errors. As can be observed, by introducing the source sentence and its translation, we are able to predict the MQM-like score of the MT output and highlight the translation errors, along with different shades of colours, according to how severe the errors are.

Our proprietary models are trained on top of a large, multilingual, pre-trained language model which caters to over 100 languages. This allows us to reliably use a single, general-purpose QE system in multiple settings across multiple language pairs. Our core, general-purpose model is trained using data generated through Human-MQM annotations, produced by professional linguists with proven experience in translation errors annotations. The MQM-QE is trained to predict translation errors in accordance with the MQM typology proposed in this study, along with the right severity. This allows us, to some extent, to use QE as an automated proxy to human annotation.

4.1. Human-MQM and MQM-QE correlation

In order to test the assumption previously stated in Section 4, we conducted an experiment with the set of East Asian Languages proposed in this study. First, we selected a corpus made of Customer Support chat content, with data coming from different clients and the following number of target words, according to the data available at the time of the experiment:

Language Pair	Number of Chat conversations	Number of Target Words
EN_JA	28	2,200
EN_ZH-TW	38	4,800
EN_KO	40	7,400
EN_ZH-CN	38	9,800

Table 8. Size of the corpora used for experiment purposes

The data were then translated with Unbabel's proprietary machine translation engines, which are transformer-based models (Vaswani *et al.* 2017) trained with the Marian toolkit (Junczys-Dowmunt *et al.* 2018). As mentioned in Gonçalves *et al.* (2022: 6), these models undergo varying levels of domain adaptation that mainly depend on the language pair and the customer. The chat messages translated in the context of this experiment were translated by using engines fine-tuned to tens to hundreds of thousands of parallel sentences of Unbabel's proprietary chat content, specific to a single client. Finally, after the machine translation, we performed two different types of evaluation: (1) human evaluation by applying the East Asian Languages MQM module proposed in this study; and (2) automatic MQM predictions by running the machine-translated data through the MQM-QE model outlined in Section 4. As for the human evaluation, this was performed by using as annotators professional linguists with previous experience in translation errors annotations. It is also important to mention that this pool of annotators was trained with the guidelines produced in this effort.

After the annotation process, we calculated the Human-MQM scores for each language pair by applying the following formula:

$$MQM = 100 - \frac{SUM((1 \times MINORS) + (5 \times MAJORS) + (10 \times CRITICALS))}{\#Words} \times 100$$

in which, to calculate the final score, each error is multiplied by the value of its severity to generate penalty points which are then summed up and divided by the number of words in the translation to obtain the final score (Lommel 2018b: 121-122).

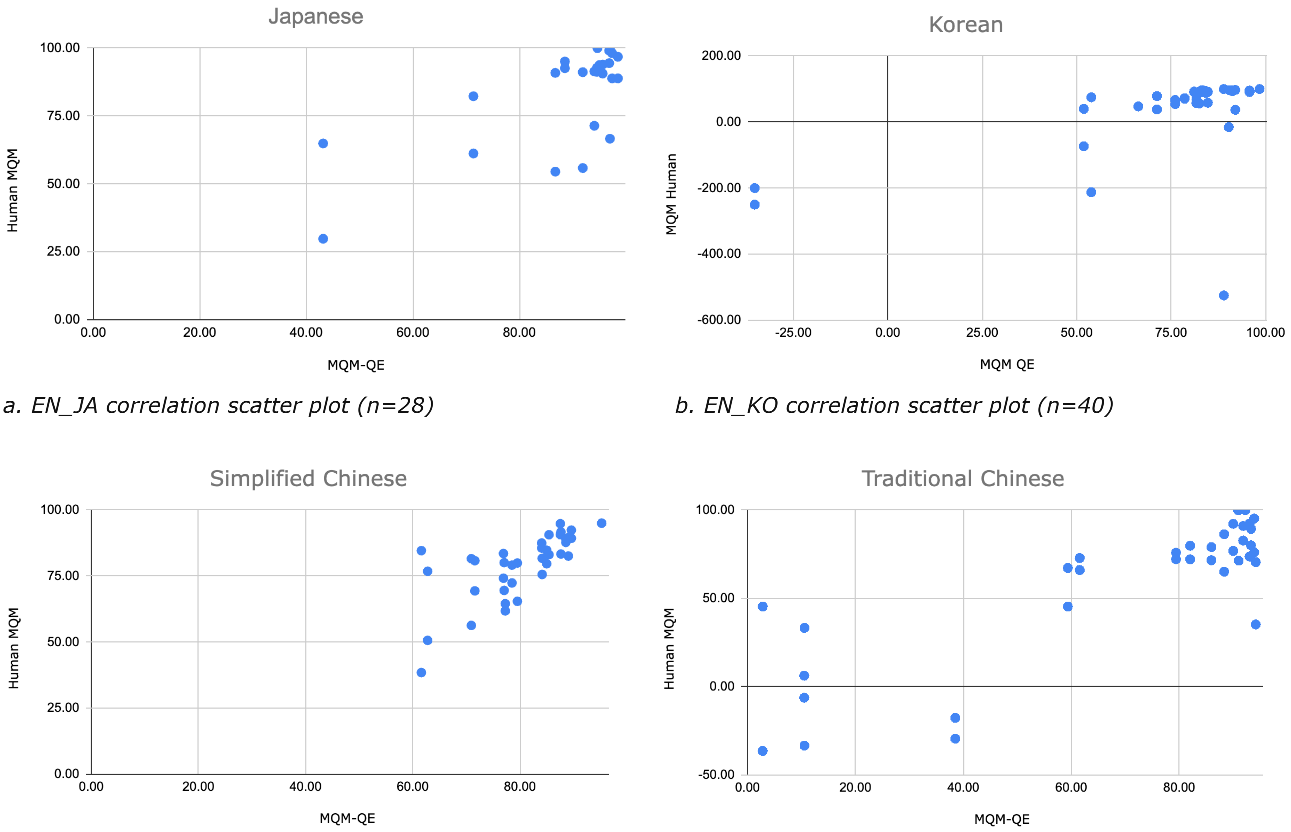
Finally, we ran the same set of jobs through the MQM-QE model here proposed, with the following results:

Language Pair	Avg. Human-MQM	Avg. MQM-QE Score
EN_KO	35.89	76.22
EN_ZH-TW	61.67	71.03
EN_ZH-CN	79.44	80.67
EN_JA	84.32	88.97

Table 9. Human-MQM and MQM-QE scores per Language Pair

Table 9 shows the results of the human annotations and the predictions provided by the MQM-QE model. One interesting example to comment on is the case of the LP EN_KO, where the average Human-MQM is 35.89 and the average score of the MQM-QE prediction is 76.22. As mentioned in Section 4, the MQM-QE engine used in this experiment is trained with proprietary Human-MQM data, where we can observe a variance in terms of annotations that can affect the final quality of the annotated data. This could be one of the reasons why it appears that the MQM-QE model is more *optimistic* in terms of final predictions, hence the difference that can be observed in Table 9 with the gold annotated data produced in the context of this experiment. In other words, in the case of EN_KO, the MQM-QE model is reproducing the *optimistic bias* that is present in the Human-MQM training data.

Finally, we measured the Pearson's *r* correlation coefficient and the *p*-value at the document level between the human-generated MQM scores - produced by using the proposed East Asian Languages MQM module - and the MQM-QE prediction, also at the document level. Figure 3 demonstrates the correlation obtained with each LP.



a. EN_JA correlation scatter plot (n=28)

b. EN_KO correlation scatter plot (n=40)

c. EN_ZH-CN correlation scatter plot (n=38)

d. EN_ZH-TW correlation scatter plot (n=38)

Figure 3. Pearson *r*’s correlation coefficients per LP

Language Pair	**Pearson’s *r***	***p*-value**
EN_KO	0.54	0.03
EN_JA	0.69	0.14
EN_ZH-CN	0.72	0.31
EN_ZH-TW	0.84	0.13

Table 10. Pearson product-moment correlation coefficient (*r*) and *p*-value per Language Pair

In Table 10 we can observe a positive correlation between the Human-MQM and the score produced by MQM-QE through the Pearson's *r*. Although one LP, EN_KO, has a result with a p-value of p<0.05, we cannot generalise this as the other tests fall short of this value. Nevertheless, the positive results of EN_KO leads us to conclude that the MQM-QE model is promising, paving the way to improve the way annotations are conducted in order to improve as well the MQM-QE model. We hope that further refinements will provide greater clarity concerning significance[14].

5. Conclusions and Future Work

The aim of this study was to propose an MQM annotation module suitable for East Asian Languages that is compliant with the MQM Framework (Lommel *et al.* 2014) and can be integrated into translation quality evaluation workflows at scale. We compared the IAA scores obtained with the module here proposed with the evaluation methodology by Ye and Toral (2020) and an MQM-compliant Error Typology used in a business setting, the Unbabel Error Typology. The IAA scores obtained with our typology were not overly superior to those corresponding to the other two typologies, but a close analysis of the annotations showed that the specific issue types and guidelines adapted for the set of languages we analysed had a positive impact in the agreement between annotators in relation to important issue types, such as those concerning function words. This is valuable not only from a linguistic perspective, as it improves the accuracy and correctness of the annotations, but also in terms of automation processes, since the increase in reliability of the annotations for these languages allows a more precise training of the automation models. Moreover, although the automatic metrics did not pass the p-test, except for the English to Korean translation direction, the present MQM-QE model trained with proprietary MQM annotated data shows promising correlation with the human annotations from our experiments. We believe that revising the annotation module and further training the annotators with basis on the shortcomings observed during the experiments on this paper can result in valuable annotation data that, if used for re-training and refining the MQM-QE model can lead to increasingly better results which will, in the future, allow accurate automatic annotations for these languages not only in terms of issues types but also spans and severities.

In this study, we propose a reference-free evaluation methodology and a way of automating Translation Quality Workflows through Quality Estimation technologies, by proposing a general-purpose MQM-QE model, trained with proprietary MQM annotated data and demonstrated that the MQM-QE model is able to predict quality scores that show a fairly high and positive correlation with Human-MQM. This poses interesting research questions and outlines new lines of work for the future to improve the quality predictions of the MQM-QE model, such as the production of a golden set of MQM-annotated data for East Asian Languages. These annotations will be then used to rescale the MQM-QE model in order to align with a much more realistic view of the quality of the translations, closer to the gold standard. Finally, a natural extension of the MQM-QE model will be to predict the correct error type, along with its associated span and severity.

6. Limitations

It is important to recognise the limitations of our experiment, which were kept in consideration while evaluating the annotation results, particularly the IAA scores, and are points that should be corrected upon further work regarding this topic.

Firstly, the fact that parent node selection was disabled for all typologies must be discussed. This is due to the fact that, as pointed out in Section 3.1.3, it negatively affected the IAA results obtained with the typology proposed by Ye and Toral (2020), which was designed to allow the selection of parent nodes. Although we do believe there is value from a linguistic analysis perspective in comparing translation error annotation typologies on a fine-grained level, there was also a limitation of the annotation tool used in these experiments, which only allows the selection of end nodes.

The second limitation of our investigation which must be addressed is the familiarity of the annotators with the annotation typologies. This is a factor that also had a negative impact on the results obtained with Ye and Toral's typology, which was different from the other two typologies that had structures familiar to the annotators: the Unbabel Error Typology because it had been used consistently by these annotators before and the East Asian Languages MQM module because it was built with a very similar structure to the former.

In light of these limitations and in an effort to present a more fair analysis of the typologies, even though we present the IAA scores, we also discuss key issue types for each error typology, regardless of IAA results, which were considered to be advantages or limitations for each typology. We recognise this does not completely eliminate the degree of direct comparisons in our overall results, which makes it relevant to present these limitations on this section.

Notes

1. Unbabel is a Portuguese software company founded in 2013 that focuses on machine translation applied to the Customer Support domain.

2. https://cordis.europa.eu/project/id/296347 (consulted 11.08.2023)

3. In this article the languages mentioned will be referred to in tables by the following codes: Japanese (JA), Korean (KO), Traditional Chinese (ZH-TW) and Simplified Chinese (ZH-CN).

4. As stated in Cabeça *et al.* (2023: 455), MQM-QE "is a system fine-tuned on Unbabel's proprietary MQM annotation data, and is designed to predict pure MQM scores with high precision".

[5] https://themqm.info/typology/ (consulted 25.11.2022)

[6] https://themqm.org/ (consulted 25.11.2022)

[7] The Unbabel Error Typology (v2) which we refer to in this paper was the error typology used for annotation within the Portuguese software company Unbabel.

[8] The general error typology referred to is the Unbabel Error Typology previously mentioned in Section 2.

[9] The annotations performed with the typology proposed in this study, as well as with the other two which are mentioned throughout this article, are carried out using three severities: Minor, Major and Critical.

[10] In total, the annotations evaluated in this phase for Japanese, Simplified Chinese and Traditional Chinese were obtained from five, three and two annotators, respectively. Although there was an attempt to always analyse data from more than one annotator, this was not possible for Korean, where the results correspond to data from one single annotator.

[11] Full version: https://photos.app.goo.gl/TVRQMyk1XAg5orAP6 and https://photos.app.goo.gl/nBJhsc5DXw1Nx5Sb9

[12] The IAA scores were computed on annotations done for different spans at document level.

[13] The MQM-QE model used in this paper is a proprietary system of a MT business and its architecture and training processes are confidential and cannot be disclosed.

[14] The authors acknowledge and are very grateful for the reviewer's suggestion for rephrasing this paragraph.

Bibliography

- **Amidei, Jacopo, Paul Piwek and Allistair Willis** (2019). "Agreement is overrated: A plea for correlation to assess human evaluation reliability." Kees van Deemter, Chenghua Lin and Hiroya Takamura (eds) (2019). *Proceedings of the 12th International Conference on Natural Language Generation*. Tokyo: Association for Computational Linguistics, 344-354.
- **Cabeça, Mariana, Marianna Buchicchio, Madalena Gonçalves, Christine Maroti, João Godinho, Pedro Coelho, Helena Monizand Alon Lavie** (2023). "Quality Fit for Purpose: Building Business Critical Errors Test Suites." Nurminen *et al.* (eds) (2023). *Proceedings of the 24th Annual Conference of the European Association for Machine Translation.* Tampere: European Association for Machine Translation, 451-460.
- **Cohen, Jacob** (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). New York: Lawrence Erlbaum Associates.
- **Freedman, David, Robert Pisaniand Roger Purves** (2007). *Statistics* (international student edition). New York: W. W. Norton & Company.
- **Gonçalves, Madalena, Marianna Buchicchio, Craig Stewart, Helena Moniz and Alon Lavie** (2022). "Agent and User-Generated Content and its Impact on Customer Support MT." Moniz *et al.* (eds) (2022). *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. Ghent: European Association for Machine Translation, 201–210.
- **Graham, Yvette, Timothy Baldwin, Alistair Moffat and Justin Zobel** (2013). "Continuous Measurement Scales in Human Evaluation of Machine Translation." Antonio Pareja-Lora *et al*. (eds) (2013). *Proceedings of the 7th Linguistic Annotation Workshop*

& Interoperability with Discourse. Sofia: Association for Computational Linguistics, 33–41.

- **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++." arXiv:1804.00344.
- **Kepler, Fábio, Jonay Trénous, Marcos Treviso, Miguel Vera and André F. T. Martins** (2019). "OpenKiwi: An Open Source Framework for Quality Estimation." arXiv:1902.08646
- **Koehn, Philipp and Christof Monz** (2006). "Manual and Automatic Evaluation of Machine Translation between European Languages." Phillip Koehn and Christof Monz (eds) (2006). *Proceedings on the Workshop on Statistical Machine Translation*. New York City: Association for Computational Linguistics, 102–121.
- **Lommel, Arle** (2018a). *Multidimensional Quality Metrics (MQM) Issue Types: DRAFT 2018-10-04*. W3C Community & Business Groups. https://www.w3.org/community/mqmcg/2018/10/04/draft-2018-10-04/ (consulted 19.11.2022).
- **Lommel, Arle** (2018b). "Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies." Joss Moorkens *et al.* (eds) (2018). *Translation Quality Assessment (Vol. 1).* Springer International Publishing, 109-127.
- **Lommel, Arle, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis and Hans Uszkoreit** (2014a). "Using a new analytic measure for the annotation and analysis of MT errors on real data." Mauro Cettolo *et al.* (eds) (2014). *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*. Dubrovnik: European Association for Machine Translation, 165–172.
- **Lommel, Arle, Hans Uszkoreit and Aljoscha Burchardt** (2014b). "Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics." *Tradumàtica: Tecnologies de La Traducció 12*, 455–463.
- **Lommel, Arle, Maja Popović and Aljoscha Burchardt** (2014c). "Assessing Inter-Annotator Agreement for Translation Error Annotation." Nicoletta Calzolari *et al.* (eds) (2014). *Proceedings of the 9th International Conference on Language Resources and Evaluation*. Reykjavik: European Language Resources Association, 1–8.
- **Lüdeling, Anke and Hagen Hirschmann** (2015). "Error annotation systems." Sylviane Granger *et al.* (eds). *The Cambridge Handbook of Learner Corpus Research (1st ed).* Cambridge University Press, 135-158.
- **Ma, Qingsong, Yvette Graham , Shugen Wang and Qun Liu** (2017). "Blend: A Novel Combined MT Metric Based on Direct Assessment — CASICT-DCU submission to WMT17 Metrics Task." Ondřej Bojar *et al.* (eds) (2017). *Proceedings of the Second Conference on Machine Translation*. Copenhagen: Association for Computational Linguistics, 598–603.
- **Macháček, Matouš and Ondřej Bojar** (2015). "Evaluating Machine Translation Quality Using Short Segments Annotations." *The Prague Bulletin of Mathematical Linguistics 103*(1), 85–110.
- **"MQM Core Typology."** https://themqm.info/typology/ (consulted 19.11.2022).
- **"MQM (Multidimensional Quality Metrics)."** https://themqm.org/ (consulted 25.11.2022).
- **Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu** (2002). "BLEU: A method for automatic evaluation of machine translation." Pierre Isabelle *et al.* (eds)

(2002). *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* Philadelphia: Association for Computational Linguistics, 311-318.

- **Popović, Maja and Mihael Arčan** (2016). "PE2rr Corpus: Manual Error Annotation of Automatically Pre-annotated MT Post-edits." Nicoletta Calzolari *et al.* (eds) (2016). *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: European Language Resources Association (ELRA), 27–32.
- **Rei, Ricardo, Craig Stewart, Ana C Farinha and Alon Lavie** (2020). "COMET: A Neural Framework for MT Evaluation." Bonnie Webber *et al.* (eds) (2020). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2685–2702.
- **Snover, Matthew, Nitin Madnani, Bonnie Dorr and Richard Schwartz** (2009). "Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric." Chris Callison-Burch *et al.* (eds) (2009). *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens: Association for Computational Linguistics, 259–268.
- **Specia, Lucia, Carolina Scarton and Gustavo Henrique Paetzold** (2018). *Quality Estimation for Machine Translation*. Springer International Publishing.
- **Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin** (2017). "Attention Is All You Need." arXiv:1706.03762
- **Vilar, David, Jia Xu, Luis Fernando d'Haro and Hermann Ney** (2006). "Error Analysis of Statistical Machine Translation Output." Nicoletta Calzolari *et al.* (eds) (2006). *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa: European Language Resources Association, 697-702.
- **Ye, Yuying and Antonio Toral** (2020). "Fine-grained Human Evaluation of Transformer and Recurrent Approaches to Neural Machine Translation for English-to-Chinese." André Martins *et al.* (eds) (2020). *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisbon: European Association for Machine Translation, 125-134.

Data availability statement

The data relevant to this research is not publicly available due to confidentiality reasons.

Biographies

Beatriz Silva is an AI Quality Analyst at Unbabel. She obtained her first BA in Asian Studies (2015) from the University of Lisbon, Portugal, and a second BA in International Chinese Language Education (2021) from the Macao Polytechnic University, China. In 2022 she received her Master's degree in Translation Studies from the University of Lisbon.

ORCID: https://orcid.org/0000-0003-1129-3331
E-mail: beatriz.silva@unbabel.com

Marianna Buchicchio is a Senior Manager, AI Quality at Unbabel. She has an MA in Translation Studies with a Major in Machine Translation (University of Lisbon). Until 2022, she was a Research Collaborator in the Group for the Computation of Lexical and Grammatical Knowledge (Linguistic Center of the University of Lisbon). She joined Unbabel in 2017, working on Quality Assurance, Human Post-Edition, MT Evaluation, Quality Technologies and NLP.

ORCID: https://orcid.org/0000-0002-2667-7867
E-mail: marianna@unbabel.com

Daan van Stigt is a Research Scientist at Unbabel. He holds a MSc degree in Logic from the Institute for Logic, Language and Computation at the University of Amsterdam, where he specialised in Machine Learning and Natural Language Processing. At Unbabel, Daan works on Quality Estimation for Machine Translation.

ORCID: https://orcid.org/0000-0001-5887-3208
E-mail:daan.stigt@unbabel.com

Craig Stewart is a Senior AI Research Manager at Phrase and was previously a Senior AI Research Manager and Team Lead in the Translation Quality Technologies Team at Unbabel. He is a specialist in Translation Evaluation and was a primary architect of COMET, the current state-of-the-art in Machine Translation automated metrics. Prior to working at Unbabel he completed an MSc in Language Technologies at Carnegie Mellon University.

ORCID: https://orcid.org/0000-0003-2649-8874

E-mail: craig.stewart@phrase.com

Helena Moniz is the President of the European Association for Machine Translation and an Assistant Professor at the School of Arts and Humanities, University of Lisbon, where she teaches Computational Linguistics, Computer Assisted Translation, and Machine Translation Systems and Post-editing. Since 2015, she is also the PI of a bilateral project with INESC-ID/Unbabel. She was responsible for the creation of the Linguistic Quality Assurance processes developed at Unbabel.

ORCID: https://orcid.org/0000-0003-0900-6938
E-mail: helena@unbabel.com

Alon Lavie is the VP of AI Research at Phrase and was until recently the VP of Language Technologies at Unbabel, where he led the development of Translation Quality Technologies. He is a Consulting Professor at the Language Technologies Institute at Carnegie Mellon University. He served as President of IAMT (2013-2015) and AMTA (2008-2012). He is a member of ACL, where he was president of SIGParse – ACL's special interest group on parsing (2008-2013).

ORCID: https://orcid.org/0000-0001-9934-6519
E-mail: alon@cmu.edu