



Universidade de Lisboa
Faculdade de Letras

Translation Error Annotation: Building an Annotation Module for East Asian Languages

Mestrado em Tradução

Beatriz Barrote Silva

2022

Relatório de estágio especialmente elaborado para a obtenção do grau de Mestre, orientado pela Professora Doutora Helena Gorete Silva Moniz e pela Mestre Marianna Buchicchio

Acknowledgements

Às minhas orientadoras, a Professora Doutora Helena Moniz e a Mestre Marianna Buchicchio, que tornaram este trabalho possível e me apoiaram durante todo o processo.

À Marianna, por ter feito tanto por mim, por ser incansável e por me ter feito sentir à vontade desde o primeiro dia.

À Professora Helena, por ter um entusiasmo contagiante pelo meu trabalho e por puxar sempre um pouco mais por mim.

À Unbabel e às equipas com que trabalhei, por me terem recebido tão bem e por me terem proporcionado uma experiência de estágio além de todas as expectativas.

Aos meus pais, por sempre me incentivarem a tentar coisas novas que mudaram a minha vida e por me ensinarem que dar o nosso melhor vale a pena.

Às minhas três irmãs, sem as quais eu não me imagino e de quem tenho sempre saudades, por fazerem de mim uma pessoa melhor.

Aos meus avós, por serem os melhores avós do mundo mesmo quando eu não sou a melhor neta e por cuidarem sempre de mim, independentemente da distância.

Às minhas amigas, as que estão perto e as que estão longe, por me ouvirem em tempos de desespero, por me fazerem rir todos os dias e por serem mais importantes do que imaginam.

Index

Abstract	7
Sumário	8
List of Tables	12
List of Figures	17
1) Introduction	18
2) Host Characterization - Unbabel	20
2.1) Unbabel and Translation Workflows	21
2.2) Quality Processes at Unbabel	24
2.2.1) Translation Errors Annotation	25
2.2.2) Post-editors' Evaluation	26
2.2.3) Automatic Quality Metrics	27
2.2.4) Tools	28
3) State of the Art	29
3.1) Machine Translation	29
3.2) Quality Processes	32
3.2.1) Manual Quality Metrics	32
3.2.2) Automatic Quality Metrics	35
3.3) East Asian Languages in Machine Translation	36
3.4) Error Typologies for Annotation	38
3.4.1) Unbabel's Error Typology	39
3.4.2) Asian Languages Focused Typology	40
4) Methodology	43
4.1) Objectives	44
4.2) Internship Tasks	45
4.2.1) Preprocessing of Asian Languages	46

4.2.2) Quality Annotation for Lingo24 Integration into Unbabel	46
4.3) Methodology for Error Annotation for East Asian Languages	47
4.3.1) Annotation Challenges	48
4.3.1.1) Missing Categories	62
4.3.1.2) Tests on Usability	71
4.3.2) The East Asian Languages Annotation Module for the Unbabel Quality Framework	72
4.3.2.1) Guidelines	82
4.3.2.2) Annotator Training	88
5) Results and Discussion	92
5.1) Unbabel Error Typology	93
5.1.1) Inter-annotator Agreement	93
5.1.2) Annotation Analysis	97
5.1.3) Annotators' feedback	105
5.2) Ye & Toral's (2020) Proposal	109
5.2.1) Inter-annotator Agreement	109
5.2.2) Annotation Analysis	113
5.2.3) Annotators' Feedback	120
5.3) East Asian Languages Annotation Module for the Unbabel Quality Framework	123
5.3.1) Inter-annotator Agreement	123
5.3.2) Annotation Analysis	128
5.3.3) Annotator's Feedback	141
6) Conclusions and Future Work	143
7) Bibliography	147
8) Annexes	154

Índice

Abstract	7
Sumário	8
Lista de Tabelas	12
Lista de Figuras	17
1) Introdução	18
2) Caracterização da Entidade de Acolhimento - Unbabel	20
2.1) A Unbabel e Processos de Tradução	21
2.2) Processos de Qualidade da Unbabel	24
2.2.1) Anotação de Erros de Tradução	25
2.2.2) Avaliação de pós-editores	26
2.2.3) Métricas de Qualidade Automáticas	27
2.2.4) Ferramentas	28
3) Estado da Arte	29
3.1) Tradução Automática	29
3.2) Processos de Qualidade	32
3.2.1) Métricas de Qualidade Manuais	32
3.2.2) Métricas de Qualidade Automáticas	35
3.3) Tradução Automática para Línguas da Ásia Oriental	36
3.4) Tipologias para Anotação de Erros	38
3.4.1) <i>Unbabel's Error Typology</i>	39
3.4.2) Tipologia focada em Línguas Asiáticas	40
4) Metodologia	43
4.1) Objetivos	44
4.2) Tarefas do Estágio	45
4.2.1) Pré-processamento de Línguas Asiáticas	46

4.2.2) Anotações de Qualidade para a Integração da Lingo24 com a Unbabel	46
4.3) Metodologia para Anotação de Erros para Línguas da Ásia Oriental	47
4.3.1) Desafios de Anotação	48
4.3.1.1) Categorias em Falta	62
4.3.1.2) Testes de Usabilidade	71
4.3.2) <i>The East Asian Languages Annotation Module for the Unbabel Quality Framework</i>	72
4.3.2.1) Diretrizes	82
4.3.2.2) Treino de Anotadores	88
5) Discussão de Resultados	92
5.1) <i>Unbabel Error Typology</i>	93
5.1.1) Concordância entre Anotadores	93
5.1.2) Análise de Anotações	97
5.1.3) <i>Feedback</i> dos Anotadores	105
5.2) A Proposta de Ye e Toral	109
5.2.1) Concordância entre Anotadores	109
5.2.2) Análise de Anotações	113
5.2.3) <i>Feedback</i> dos Anotadores	120
5.3) <i>East Asian Languages Annotation Module for the Unbabel Quality Framework</i>	123
5.3.1) Concordância entre Anotadores	123
5.3.2) Análise de Anotações	128
5.3.3) <i>Feedback</i> dos Anotadores	141
6) Conclusões e trabalho futuro	143
7) Bibliografia	147
8) Anexos	154

Abstract

In this thesis it is proposed an annotation module to be applied in the context of Machine Translation (MT) concerning the East Asian languages of Japanese, Korean and Mandarin for the purpose of assessing MT output quality through annotation. The annotation module was created based on a data-driven analysis over Customer Support content in these languages previously annotated with the Unbabel Error Typology, which is a general typology in the sense that it is not conceived for any specific groups of languages. As such, this work also explores how applying translation error typologies inadequate to certain languages or content types can have an impact on how annotation reflects the quality of a translation.

For the purpose of testing the effectiveness of the proposed annotation module, an annotation experiment for the languages under analysis was conducted. This experiment consisted of, for each language, annotating the same content using three different error typologies: the Unbabel Error Typology, the MQM-compliant error taxonomy for the translation direction of English to Chinese proposed by Ye and Toral (2020) and the annotation module proposed on this thesis. Furthermore, each dataset was annotated by two annotators. This allowed a comparison of Inter-annotator agreement (IAA) scores, which constitutes an important metric in terms of evaluating the effectiveness of an error typology.

In light of this, each of the tested typologies was analyzed based on the obtained IAA scores and a further in-depth analysis of concrete annotations which lead to an understanding over their strengths and limitations.

With this work it was possible to demonstrate that, if on one hand using error typologies inadequate for the content annotated has a negative impact on the quality of said annotations, on the other hand applying an error typology specific to the content being annotated can result in more consistent annotations.

Keywords: Multidimensional Quality Metrics Framework; East Asian Languages; Machine Translation Evaluation; Inter-Annotator Agreement; East Asian Languages Evaluation

Sumário

O trabalho desenvolvido no âmbito desta tese teve como objetivo principal a criação de um módulo de anotação para erros de tradução no contexto da Tradução Automática (TA) que fosse aplicável a Japonês, Coreano e Mandarim e compatível com o *Multidimensional Quality Metrics (MQM) framework* (Lommel et al., 2014). Este módulo foi criado com base numa análise de dados reais sobre traduções previamente anotadas dentro da empresa Unbabel seguindo uma tipologia geral concebida para anotação de vários pares linguísticos sem foco em grupos de línguas específicos. Ao mesmo tempo que permitiu verificar as consequências de anotar erros com uma tipologia pouco adequada à língua ou ao conteúdo traduzido, esta análise constituiu um ponto de partida importante para a criação do módulo de anotação proposto nesta tese.

A **Secção 2** desta tese concentrou-se em apresentar a Unbabel como instituição e os processos de qualidade em vigor dentro da empresa. A **Secção 3** focou-se em apresentar o estado da arte em TA e processos de qualidade, com atenção especial às línguas sob análise nesta tese, bem como as tipologias de anotação de erros de tradução utilizadas para comparação de resultados.

A análise dos dados disponíveis, descrita na **Secção 4**, foi feita em duas fases principais. Na primeira fase foi analisado um conjunto de 342 segmentos correspondentes ao par linguístico Inglês-Chinês (Simplificado), previamente anotados com a *Unbabel Error Typology*, a tipologia para anotação de erros de tradução utilizada para todos os pares linguísticos até junho de 2022. Esta análise demonstrou que uma percentagem significativa dos erros cometidos durante o processo de anotação podiam ser atribuídos não só à falta de clareza das diretrizes de anotação relativamente a características específicas presentes neste par linguístico como também à falta de alguns tipos de erros na tipologia. Na segunda fase de análise de dados foi possível confirmar e fundamentar a existência destes problemas. Nesta fase foi analisada uma amostra de dados mais abrangente que incluiu quatro pares linguísticos: Inglês-Japonês, Inglês-Coreano, Inglês-Chinês (Simplificado) e Inglês-Chinês (Tradicional). Para cada par linguístico foi analisado um total de cerca de 570 a 1900 segmentos e, com a exceção de Inglês-Coreano, todos os dados correspondiam às anotações de mais de um anotador. Esta análise permitiu concluir que os anotadores de todos os pares linguísticos mencionados cometeram vários erros, em especial no

processo de escolha da categoria certa para cada erro de tradução mas também relativamente à seleção dos erros e atribuição da severidade certa a cada um.

Através dos dados analisados foi possível determinar que tipos de erros seria necessário incluir numa tipologia de anotação de erros de tradução adaptada às línguas mencionadas e que tipo de instruções deveriam ser clarificadas nas diretrizes de anotação. Deste modo, após a conclusão da segunda fase de análise de dados foi possível começar a criar o módulo de anotação proposto nesta tese, denominado *East Asian Languages Annotation Module for the Unbabel Quality Framework*.

O *East Asian Languages Annotation Module for the Unbabel Quality Framework* foi criado à imagem da *Unbabel Error Typology* e adaptado às características da nova versão que entrou em vigor na empresa em junho de 2022. No entanto, devido ao facto de ser um módulo de anotação adaptado às línguas asiáticas previamente mencionadas, várias categorias de erros existentes na *Unbabel Error Typology* foram removidos devido a corresponderem a componentes linguísticos que não fazem parte das línguas em questão. Do mesmo modo, foi adicionado um total de cinco novos tipos de erros ao módulo com base no que foi julgado necessário durante a fase de análise de dados. A versão final do *East Asian Languages Annotation Module for the Unbabel Quality Framework* conta com um total de 39 tipos de erros, em contraste com os 47 que fazem parte da *Unbabel Error Typology*. De forma complementar à criação do módulo de anotação foram também elaboradas diretrizes específicas para o módulo que, para além da definição de cada tipo de erro com exemplos, incluem também uma secção dedicada a casos difíceis (*Tricky Cases*) e esquemas (*Decision Trees*) para auxiliar na escolha da severidade e tipo de erro adequado para cada caso.

Após a criação do módulo de anotação foi necessário testar se o mesmo pode ser aplicado com sucesso. Para este fim foi levado a cabo um estudo de comparação entre o *East Asian Languages Annotation Module for the Unbabel Quality Framework* e duas outras tipologias, descrito na **Secção 5**. Assim, foram conduzidas três fases de anotação com cerca de um mês de intervalo entre cada. Para cada tipologia dois anotadores por par linguístico anotaram entre 1100 e 4900 palavras cada um e, de modo a obter uma comparação precisa, dentro de cada par linguístico o conteúdo anotado com cada tipologia manteve-se o mesmo.

A primeira fase de anotações foi efetuada utilizando a *Unbabel Error Typology*. Devido ao facto de os anotadores já estarem familiarizados com esta tipologia e já possuírem as

diretrizes de anotação relativas à mesma, não foi necessário prestar apoio adicional aos anotadores nesta fase.

A segunda ronda de anotações foi levada a cabo com a tipologia para anotação de erros de tradução para o par linguístico Inglês-Mandarim proposta por Ye e Toral (2020). Para esta fase de anotação foram criadas diretrizes específicas para esta tipologia com base no trabalho desenvolvido por Ye e Toral (2020) de modo a facilitar o processo de anotação. É importante referir que, apesar de esta tipologia ter sido criada para anotação de erros de tradução para o par linguístico Inglês-Mandarim, durante a fase de teste das tipologias esta foi utilizada para anotar todos os quatro pares linguísticos a serem analisados. Além disso, devido ao facto de ser uma tipologia nova, durante esta fase foi mantida a comunicação com os anotadores para esclarecimento de dúvidas. É necessário salientar que esta tipologia também foi importante na criação do *East Asian Languages Annotation Module* devido ao facto de conter tipos de erros específicos em relação à anotação do par linguístico para o qual foi criada e que serviram de base para novos tipos de erros propostos no módulo de anotação.

A terceira e última fase de anotação foi feita com o *East Asian Languages Annotation Module for the Unbabel Quality Framework* proposto nesta tese. Nesta fase foram fornecidas aos anotadores as diretrizes que foram criadas de forma complementar ao módulo e, tal como durante a segunda fase, foi dada aos anotadores a possibilidade de comunicar as suas dúvidas.

Os resultados das três fases de anotação descritas acima foram analisados da perspetiva do nível de acordo entre os anotadores, medido através da metodologia de *Inter-annotator Agreement* (IAA), em contraste com os valores equivalentes da métrica manual de qualidade MQM (Lommel et al., 2014), bem como através de uma análise detalhada das anotações de ambos anotadores para todos os pares linguísticos. No contexto da testagem de tipologias de anotação de erros de tradução uma análise dos valores de IAA obtidos, pois um elevado nível de concordância entre os anotadores reflete a clareza de uma tipologia. Adicionalmente, a análise detalhada das anotações em consonância com os valores de IAA permite avaliar que fatores influenciam a flutuação dos mesmos. Adicionalmente, o *feedback* que os anotadores forneceram em relação a cada tipologia também foi alvo de reflexão em contraste com os resultados obtidos. Deste modo, com a combinação de todos estes dados foi possível determinar os pontos fortes e as fraquezas de cada tipologia bem como entender que direção deverá seguir o trabalho futuro

em torno do *East Asian Languages Annotation Module for the Unbabel Quality Framework* em termos do seu aperfeiçoamento.

Com este trabalho foi possível demonstrar o impacto negativo de utilizar uma tipologia de erros pouco adequada ao conteúdo a ser anotado bem como provar que, por outro lado, uma tipologia criada para a anotação de um grupo específico de línguas pode melhorar a consistência das anotações relativas a componentes linguísticos próprios das línguas para as quais a tipologia é direcionada.

Palavras-chave: *Multidimensional Quality Metrics Framework*; Línguas da Ásia Oriental; Avaliação de Tradução Automática; Concordância entre Anotadores; Avaliação de Línguas da Ásia Oriental

List of Tables

Table 1 Percentage of annotation errors per LP	58
Table 2 Wrong span annotation in Japanese	61
Table 3 New issue types related to particles	62
Table 4 Definition of <i>Omitted Particle</i> errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework	65
Table 5 Definition of <i>Wrong Particle</i> errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework	66
Table 6 New issue types related to classifiers	66
Table 7 Classifiers in Chinese	67
Table 8 Classifier issue type in the typology proposed by Ye and Toral (2020)	67
Table 9 Definition of <i>Omitted Classifier</i> errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework	68
Table 10 Definition of <i>Wrong Classifier</i> errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework	69
Table 11 New issue type for transliteration	69
Table 12 Definition of <i>Transliteration</i> errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework	70
Table 13 Definition of <i>Extraneous</i> errors in the guidelines created for the typology proposed by Ye and Toral (2020)	79
Table 14 Comparison of issue types between the typology proposed by Ye and Toral (2020) and the East Asian Languages Annotation Module for the Unbabel Quality Framework	80

Table 15 Definition of <i>Tense/Mood/Aspect</i> errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework	81
Table 16 Definition of <i>Punctuation</i> errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework	83
Table 17 Tricky Cases section on annotating verbs with the East Asian Languages Annotation Module for the Unbabel Quality Framework	85
Table 18 Average batch IAA per LP with the Unbabel Error Typology	94
Table 19 Example of annotation of whitespaces in Korean with the Unbabel Error Typology	95
Table 20 Example of span selection in Simplified Chinese with the Unbabel Error Typology	96
Table 21 Average batch MQM per LP and annotator with the Unbabel Error Typology	96
Table 22 Example of EN-JA annotation of particles with the Unbabel Error Typology	98
Table 23 Example of EN-ZH_CN annotation of classifiers with the Unbabel Error Typology	99
Table 24 Example of different span selection in the annotation of the <i>Omission</i> type with the Unbabel Error Typology for JA	100
Table 25 Example of annotation of <i>Omission</i> with the Unbabel Error Typology for ZH-CN	101
Table 26 Example of annotation of <i>Punctuation</i> in Simplified Chinese with the Unbabel Error Typology	102
Table 27 Examples of annotation of <i>Mistranslation</i> with the Unbabel Error Typology across all LPs	103
Table 28 Examples of EN-KO and EN-ZH_TW annotation of <i>Register</i> with the Unbabel Error Typology	104
Table 29 Examples of mismatch of severities with the Unbabel Error Typology in Japanese and Simplified Chinese	105

Table 30 Annotators' feedback on not applicable issue types in the Unbabel Error Typology	106
Table 31 Annotators' feedback on missing issue types in the Unbabel Error Typology	107
Table 32 Average batch IAA per LP with Unbabel's and Ye and Toral's typologies	109
Table 33 Examples of issue type disagreement with Ye and Toral's typology in Traditional Chinese	111
Table 34 Average batch MQM per LP and annotator with Ye and Toral's typology	112
Table 35 Example of EN-JA annotation of omitted function words with Ye and Toral's typology	114
Table 36 Example of EN-JA annotation of function words with Ye and Toral's typology	114
Table 37 Example of annotation of EN-ZH_CN classifiers with Ye and Toral's typology	115
Table 38 Definition of <i>Classifier</i> errors in Ye and Toral's typology	116
Table 39 Example of EN-JA annotation of <i>Mistranslation</i> with Ye and Toral's typology	117
Table 40 Examples of EN-KO and EN-ZH_TW annotation using different issue types with Ye and Toral's typology	118
Table 41 Example of EN-ZH_CN annotation of <i>Terminology</i> with Ye and Toral's typology	119
Table 42 Examples of unification of issue types with Ye and Toral's typology in Japanese	120
Table 43 Annotators' feedback on missing issue types in Ye and Toral's typology	121
Table 44 Average batch IAA per LP with the Unbabel Error Typology, the typology proposed by Ye and Toral and the East Asian Languages Annotation Module for the Unbabel Quality Framework	124
Table 45 Example of EN-JA annotation of <i>Register</i> and <i>Terminology</i> with the East Asian Languages Annotation Module for the Unbabel Quality Framework	126

Table 46 Average batch MQM per LP and annotator with the East Asian Languages Annotation Module for the Unbabel Quality Framework	127
Table 47 Examples of EN-JA and EN-KO annotation of particles with the East Asian Languages Annotation Module for the Unbabel Quality Framework	129
Table 48 Example of EN-ZH_CN annotation of classifiers with the East Asian Languages Annotation Module for the Unbabel Quality Framework	130
Table 49 Example of EN-KO and EN-ZH_TW annotation of <i>Transliteration</i> with the East Asian Languages Annotation Module for the Unbabel Quality Framework	131
Table 50 Example of EN-JA annotation of <i>Mistranslation</i> with the East Asian Languages Annotation Module for the Unbabel Quality Framework	132
Table 51 Examples of EN-JA, EN-ZH_TW and EN-ZH_CN annotation of <i>Mistranslation</i> with the East Asian Languages Annotation Module for the Unbabel Quality Framework	133
Table 52 Example of EN-KO issue type disagreement with the East Asian Languages Annotation Module for the Unbabel Quality Framework	134
Table 53 Example of annotation of <i>Tense/Mood/Aspect</i> in Simplified Chinese with the East Asian Languages Annotation Module for the Unbabel Quality Framework	135
Table 54 Annotation of <i>Tense/Mood/Aspect</i> in Korean with the East Asian Languages Annotation Module for the Unbabel Quality Framework	135
Table 55 Example of EN-JA and EN-ZH_CN agreement on Span and Severity with the East Asian Languages Annotation Module for the Unbabel Quality Framework	136
Table 56 Example of EN-JA disagreement on Span and Severity with the East Asian Languages Annotation Module for the Unbabel Quality Framework	137

Table 57 Example of EN-JA issue type disagreement with the East Asian Languages Annotation Module for the Unbabel Quality Framework	138
Table 58 Definition of <i>Do not Translate</i> errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework	138
Table 59 Definition of <i>Punctuation</i> errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework	139
Table 60 Example of annotation of <i>Punctuation</i> in Simplified Chinese with the East Asian Languages Annotation Module for the Unbabel Quality Framework	140
Table 61 Examples of EN-KO and EN-ZH_CN agreement on <i>Register</i> and <i>Terminology</i> with the East Asian Languages Annotation Module for the Unbabel Quality Framework	141

List of Figures

Figure 1 Chat translation pipeline	22
Figure 2 Post-editor's assessment rating system	27
Figure 3 Unbabel Error Typology	40
Figure 4 The MQM-compliant error taxonomy for the translation direction English-Chinese (Ye and Toral, 2020)	41
Figure 5 Percentage of annotation errors in Simplified Chinese dataset	49
Figure 6 Percentage of annotation errors for all LPs	56
Figure 7 Example of annotation analysis	57
Figure 8 Example of wrong annotations in Japanese	59
Figure 9 Sentence-final particles in Japanese extracted from Makino and Tsutsui (1989:45)	64
Figure 10 East Asian Languages Annotation Module for the Unbabel Quality Framework	74
Figure 11 Decision Tree for issue type selection in the East Asian Languages Annotation Module for the Unbabel Quality Framework	87
Figure 12 Decision Tree for severity selection in the East Asian Languages Annotation Module for the Unbabel Quality Framework	88

1. Introduction

This thesis was written for the Master's degree on Translation of the School of Arts and Humanities of the University of Lisbon in the context of an internship at Unbabel, a portuguese software company which offers translation services in the Customer Support, market and product information domain.

Unbabel is a company that offers translation services based on a hybrid system that combines machine translation (MT) with human post-editing. In this context, translation quality assessment is a key factor to ensure the reliability of the services provided. One of the quality processes used at Unbabel for this end is human annotation of translation errors, which results in data that is used to continuously train the MT systems. Annotation is performed based on a list of possible errors the annotators can identify in a translation. These lists are referred to as error typologies. If on one hand an error typology that is overly extensive and detailed can overwhelm annotators and complicate the annotation process, on the other hand an error typology that is too simple may not be sufficient and may even affect annotations equally. This means that it is quite challenging to have an adequately sized error typology that fits all languages and content types equally, if possible at all with the state-of-the-art on error typologies.

The work developed during this internship at Unbabel aims to demonstrate the negative consequences of applying the same error typology independently from the language that is being annotated and to propose a solution for this limitation. The data analysis on this thesis will be focused on the East Asian languages of Japanese, Korean and both the Traditional and Simplified written variants of Chinese, which constitute a specific language set for which the typology in use at Unbabel may be inadequate due to it being conceived from the perspective of Western languages. As such, it is expected that this typology will contain several issue types that do not apply to the languages at hand and, at the same time, lack issue types that are essential to obtain accurate annotations for these languages.

In light of this, the ultimate objective of this project will be to propose an annotation module adapted to the characteristics of this specific set of languages and to measure its success through comparison with Unbabel's typology and another error typology which has been

conceived specifically for annotation in the translation direction of English to Chinese. In order to do so, the first step will be to conduct an in-depth analysis of already existing annotation data for Japanese, Korean, Traditional and Simplified Chinese within Unbabel. This analysis will serve the purpose of determining what the annotators for these languages struggle the most with, in order to learn what problems the proposed annotation module should address. Once the annotation module has been created, it will be essential to test it in order to determine what points were successfully implemented and which will need further work in future iterations of the annotation module.

The testing phase of the annotation module will be conducted through three phases of annotation, each one corresponding to a different typology. The first phase of annotation will be performed using the Unbabel Error Typology, which was in use at the company until June of 2022. In the second phase the annotators will use an MQM-compliant typology created for the translation direction of English to Chinese proposed by Ye and Toral (2020). Finally, the third and last phase of annotations will be conducted using the annotation module which will be proposed in this thesis. During this process, two annotators for each language pair will annotate a similar number of jobs and words with all three typologies, which will serve the purpose of pinpointing the strong and weak points of each typology through not only an in-depth analysis of the annotations but a comparison of the Inter-annotator Agreement (IAA) scores, thus determining whether the annotation module proposed on this thesis can make a positive contribution to the annotation process at Unbabel and be productized in a future step.

In **Section 2** of this thesis Unbabel, the host institution for this work, will be presented along with the quality processes in place at the company. Following that, **Section 3** will provide an overview of the state-of-the-art of MT and translation quality assessment processes, focusing also on the specific case of the East Asian languages at hand. **Section 4** will discuss the methodology of this project, including not only the data analysis that was carried as basis for building the proposed annotation module but also the tasks developed during the internship as well as the process of creating the annotation module for East Asian Languages. In **Section 5** the results of the annotation process using all three typologies will be presented and discussed. Finally, **Section 6** will contain the final conclusions of this work as well as a reflection on the future work that can be developed based on the work developed during this internship.

2. Host Characterization - Unbabel

Unbabel is a portuguese software company founded in 2013 that focuses on translation applied to Customer Support. With the growing need for communication between different languages also comes the increasing necessity for translation services to deal with languages we are not familiar with. However, if on one hand human translation can be very slow and expensive, on the other hand Machine Translation is still far from perfect, despite all the recent developments with neural systems. Unbabel's mission is to create a hybrid system that makes use of both of these types of translation by having a community of human post-editors reviewing machine translations, allowing for faster translations that are also good in quality.

Until January of 2022, Unbabel operated with 27 different language pairs by employing a community of freelance post-editors and annotators spread throughout the world. In addition, apart from headquarters in Lisbon, Unbabel already had offices in San Francisco, New York and Pittsburgh and hubs in London and Berlin. In December of 2021 Unbabel acquired Lingo24, a tech-enabled Language Service Provider (LSP) headquartered in Edinburgh and with offices in Cebu (Philippines) and Timișoara (Romania). With the integration of Lingo24, Unbabel was able to expand its offer significantly, as since its foundation in 2001 Lingo24 has provided services in many areas outside of Customer Support in an unlimited number of language pairs, since it collaborates with over 4,000 translators, transcreators and copywriters who are mostly based in the country of their target language.

The following sections in this chapter will present how Unbabel works. In **Section 2.1.** the translation workflows presently in place at Unbabel will be explained, particularly the sections that are relevant for the work developed during this internship, as well as a brief overview of how the integration with Lingo24 affected them. Following that, **Section 2.2.** will present the quality processes that are used at Unbabel in order to ensure quality in the services provided and continuous improvements.

2.1. Unbabel and Translation Workflows

Up until the integration of Lingo24, Unbabel's services were focused exclusively on the area of Customer Support. Therefore, translation quality standards were also defined in conformity with this purpose and the content types that were handled were also customer support oriented and included chat, tickets¹ and Frequently Asked Questions (FAQs). The processes of translation until the delivery step for each of these were represented through pipelines that contained more or less steps depending on the content type and its idiosyncratic requirements. The reason why different content types correspond to different translation pipelines is that the translation quality and speed of delivery required for each of them are different. While chat and tickets consist of one-to-one interactions, FAQs are a one-to-many type of communication, meaning that the target audience is much broader. In addition, FAQs are usually published in the official website of companies and may be a direct reflection of their image. For this reason, the quality required for this content type is much higher. However, there are also differences between chat and tickets due to the fact that the turnaround time (TAT) for chat is much shorter, since it consists of a type of real-time communication. This means that there is no time to post-edit so the expected quality for the translations is also lower.

Since the investigation developed in this thesis was conducted using chat data, it is important to look at the pipeline represented in **Figure 1** and go into some detail about it.

¹ "Tickets" is the term used to refer to emails in Customer Support jargon.

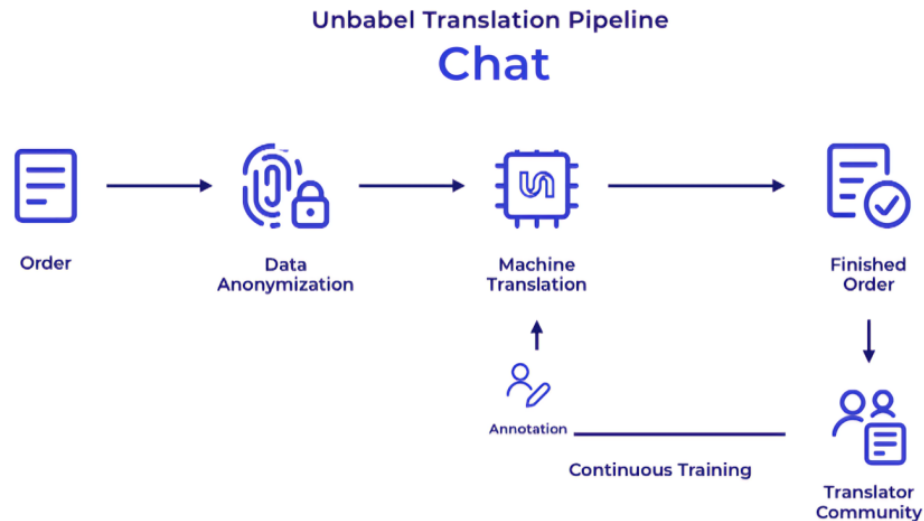


Figure 1. Chat translation pipeline

Chat is a type of communication that is meant to happen in real time. Therefore, when there is a process of translation involved, the speed of delivery is a priority. Once the order is placed by the customer through CRMs² such as Salesforce or Zendesk, the translation is generated through the process illustrated in **Figure 1**. In order to meet the time constraint requirements, the quality of these translations is assured by pre-trained models based on human post-edition rather than post-edition efforts after the translation is ready. In terms of the Unbabel translation pipeline, this means that the translation produced in the machine translation step is the final product that will be delivered to the client. It is only after this that the translation is annotated by Unbabel’s community, so that the annotation data can be used to train the machines and improve the quality of future translations.

In addition, it is also important to mention that independently from the type of content being translated, one core step of Unbabel’s translation pipeline is data anonymization. Because customer service usually deals with sensitive information³, such as names, contact information,

² As defined by Salesforce, “Customer Relationship Management (CRM) is a strategy that companies use to manage interactions with customers and potential customers. CRM helps organizations streamline processes, build customer relationships, increase sales, improve customer service, and increase profitability.” (<https://www.salesforce.com/eu/learning-centre/crm/what-is-crm/>)

³ This sensitive information is denominated PII (Personal Identifiable Information).

banking details, secret passwords and codes, in accordance with the European Union’s General Data Protection Regulation (GDPR)⁴, it is essential to have a step before translation that hides this information, in order to prevent malicious use. As such, this type of information is detected automatically and is then hidden behind specific tags, allowing only a censored version of the text to go into the translation process. After all the translation steps have been completed, the anonymized entities are automatically restituted, so that the information is present in the final product that is delivered to the client.

Since Unbabel acquired Lingo24, not only more content types were introduced, but also more translation processes that demanded a new approach to quality control. With the integration, the strategies for communicating quality at Unbabel shifted to a new scheme named Quality Framework, which divides translation quality into six different levels. The principle behind having different quality levels is that not all types of content demand the same level of translation quality and the objective is to provide the clients with translations that are within what they expect in terms of quality depending on the translated content and corresponding workflows.

The six quality levels that make up the Quality Framework are based on Multidimensional Quality Metrics (MQM) scores (Lommel et al., 2014), meaning that each level corresponds to increasingly high expected median MQM score thresholds. These levels serve the purpose of communicating to clients the quality level that can be expected for each content type, which varies according to the translation process behind it. For example, Level 0, which corresponds to lower expected median MQM scores, includes content that is the product of Machine Translation (MT) with no post-edition, such as chat translations, which were mentioned previously and are meant to be delivered almost in real-time. On the other hand, Levels 4 and 5, which correspond to almost perfect MQM, include translations that are expected to be of extremely high quality since they are performed by professional translators, which is a workflow that only became a reality at Unbabel after the integration. In the case of Level 5, since it includes “on-brand translation”, the required quality is especially high.

⁴ The GDPR is a directive regulated by the European Parliament which dictates “the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing” <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN>

Since the main object of investigation in this work is chat data, it is important to further explain what is expected in terms of quality for chat translation according to the Quality Framework. As mentioned previously in this section, due to the fact that this type of communication takes place in real-time and instant results are expected, speed is prioritized over quality and the main purpose is to provide content that is understandable in very short turn-around times. This means that the translation that reaches the end-user is the product of unedited MT outputs and, as such, it most likely will contain issues that could possibly have an impact on fluency but will, ultimately, still be understandable and get the intended message across. Due to these characteristics, only customer service chats and translation simulations are included under this level, since it is inadequate for other content types.

2.2 Quality Processes at Unbabel

Unbabel relies on a process of post-edition and annotation of Machine Translation output in order to ensure the quality of the delivered product and its continuous improvement. For this, aside from the community of post-editors and annotators in charge of improving the quality of the translation and the performance of the machines, Unbabel also works with experts, such as professional translators and linguists, that can act as annotators, terminologists, evaluators or language consultants, and several tools that work towards the same goal of ensuring quality translations that are AI based. In addition, there are also several automatic processes in place to predict translation quality and improve the content that is delivered to the end-users. As such, this section will start by explaining the manual quality metrics which are in use at Unbabel and constitute a fundamental basis for the work developed during this internship in **Section 2.2.1.**, as well as the automatic quality metrics that further ensure translation quality at Unbabel in **Section 2.2.2.**

2.2.1. Translation Errors Annotation

According to Lüdeling and Hirschmann (2015), annotation is the assignment of a category to a segment of the corpus, said category usually being drawn from a predefined or finite tagset.

In Unbabel's pipeline, annotation takes place after translations are delivered to the client. This process consists of identifying the errors in the translation in order to attribute MQM score to it and use it to continuously improve the MT output. The official definition for Multidimensional Quality Metrics (MQM) states that it is "a framework for describing and defining quality metrics used to assess the quality of translated texts and to identify specific issues in those texts"⁵. As will be explained in more detail in **Section 3.2.1.**, the MQM framework was developed by Arle Lommel, in 2014 in the scope of the QTLaunchPad, and is currently the standard for manual quality evaluation in the translation industry. One of the main characteristics of the MQM framework that is important to mention in terms of annotation at Unbabel is the fact that it was designed with the intent of being customizable, both in terms of selectable issue types and attributed severities, and is, therefore, adaptable to the objectives of its users. The typology in use at Unbabel until June of 2022, explained in more detail on **Section 3.4.1.**, was a custom typology developed with MQM at its core but built and customized with the intent to suit annotation in the Customer Support domain. After this date, that typology was replaced by a new MQM-compliant version meant to accommodate the specificities of the content brought in by Lingo24.

Annotations are useful in the process of continuous improvement of translation quality, not only because they allow the identification of errors that should be corrected, but also because they can be used to improve consistency with annotations. As Lüdeling and Hirschmann (2015) explain, the evaluation of annotations through comparison with a gold standard or between how different annotators, when provided the same dataset and guidelines, annotate the same subcorpus, or in other words, how high is the Inter-Annotator Agreement (IAA). This can show if the definition of each category is clear or if it should be improved and, ideally, contributes to a cyclic process of improvement in which after the evaluation of the annotations the guidelines are

⁵ <https://www.qt21.eu/mqm-definition/definition-2015-12-30.html>

adjusted to be used again and repeat the process until consistency is achieved (Lüdeling & Hirschmann, 2015).

At Unbabel, the annotation process is done through Unbabel's proprietary annotation tool using a specific error typology that was developed on the basis of the MQM framework. Unlike previous models, the MQM framework was built with the intention to be a flexible and customizable system for evaluating translation quality which allows different levels of granularity concerning issue types (Lommel et al., 2014). As such, the Unbabel Error Typology, which will be explained in more detail in **Section 3.4.1.**, consisted of three main categories derived from MQM's core (accuracy, fluency and style) that branch out into sixteen parent tags, thirty daughter tags and ten granddaughter tags.

2.2.2. Post-editors' Evaluation

In order to ensure certain standards of quality, Unbabel carries out pre-onboarding and continuous evaluations of the community of post-editors, since their role in the pipeline means they have an important impact both in the training of the machines and on the quality of the delivered product.

To carry out the initial testing process, Unbabel provides specific guidelines for each language and usability guidelines for the tools the community has to use to prepare them in the best way possible. After post-editors pass the testing phase, they enter the training phase and go through further evaluation in order to determine whether they are qualified to have access to paid tasks. However, in order to ensure quality is maintained, once they are already established members of the community who are in charge of paid tasks, post-editors still have to go through periodical assessments to determine the quality of their job and depending on the results they may maintain their status or be demoted from paid to training tasks once again. In order to carry out these evaluations, random tasks from each post-editor are randomly selected and then assigned to an Evaluator, who is a professional translator or linguist. The quality of editions is assessed by these evaluators and represented through a star rating system that goes from one star (bad quality) to five stars (excellent quality), as shown in **Figure 2.**



Figure 2. Post-editor’s assessment rating system

2.2.3. Automatic Quality Metrics

When handling a big volume of translations it is both very costly and time consuming to have human evaluation as the exclusive method of measuring their quality. As such, there are several automatic metrics in place at Unbabel to assess the quality of MT outputs, which reduce the need for human intervention.

One of the most important processes is Quality Estimation (QE) (Martins et al., 2017), which determines the quality of a translation. Unlike other automatic quality metrics, which will be further discussed in **Section 3.2.2.**, QE is an automatic method “for estimating the quality of a machine translation (MT) output at run-time, without the use of reference translations” (Specia et al., 2018). In other words, QE is a process that measures how good a translation is and it is a process that, at Unbabel, is done automatically by an in-house developed software named OpenKiwi, a “source framework for QE that implements the best QE systems from WMT 2015-18⁶ shared tasks, making it easy to combine and modify their key components, while experimenting under the same framework” (Kepler et al., 2019). As pointed out by the authors, Kiwi aims to solve the shortcomings of previous QE systems which were not easily reproducible due to their complexity.

Similarly to other QE models, Kiwi (Kepler et al., 2019) attributes a score to translations based on how the machine translation (MT) sentence relates to the source or by comparing the

⁶ “WMT is the main event for machine translation and machine translation research. The conference is held annually in connection with larger conferences on natural language processing.” <https://machinetranslate.org/wmt>

source text to the post-edited translation, without using any references. In terms of the translation pipeline, Kiwi determines whether an automatic translation is good enough to be delivered, avoiding post-edition efforts on translations that fall above a certain threshold of quality.

In addition to Kiwi, another important automatic process in place at Unbabel that works towards quality is COMET which, as will be mentioned later in **Section 3.2.2.**, consists of a “neural framework for training multilingual machine translation evaluation models” (Rei et al., 2020) and is used at Unbabel to determine whether a new engine is fit to be deployed for production, or in other words, to measure how good the MT system is. This metric is based on years of proprietary MQM annotations and, because of this, shows very high correlation with human judgment.

2.2.4. Tools

In addition to the proprietary tools for annotation and post-editors evaluation, used for annotating and conducting evaluation on the community of post-editors respectively, Unbabel also makes use of a number of tools and resources with the purpose of enhancing the quality of both MT and post-editing (PE) processes, such as glossaries and Unbabel’s errors identification tool - Smartcheck. Similarly to what happens with the community of post-editors, the performance of each of these tools has a significant impact on the quality of the translations and, as such, are submitted to continuous improvement and development.

This chapter focused on explaining what Unbabel is as a company and the changes that it went through in the duration of this internship that both affected it and generated new opportunities, as will be discussed later in **Section 4.2.** In addition, it also discussed the processes that Unbabel makes use of to ensure the best possible quality in the delivered product, both with or without direct human intervention.

3. State of the Art

The advancement of technology in the past century has created many opportunities to design solutions for problems we have always struggled with. As communication between different languages started to expand, the demand for translation services also started to surpass the availability of existing translators. As such, Machine Translation (MT) started being regarded as a possible solution and research in this area started growing in the first half of the twentieth century.

In this chapter a summary of the history of machine translation will be presented, followed by an analysis of the quality processes that have been developed over the years in an attempt to evaluate the quality of machine translation outputs. Following that, **Section 3.3.** will present a brief overview of the research done to date concerning machine translation applied to the four East Asian languages under discussion on this thesis. Finally, the annotation typologies that will be used for comparison with the annotation module proposed on this thesis will be presented in **Section 3.4.**

3.1. Machine Translation

The twentieth century marks the beginning of the first proposals of machines for automatic translation (Hutchins, 2003). More specifically, in 1933, two patents belonging to Georges Artsrouni and Petr Trojanskij were issued in Europe for two different machines that could be considered mechanical dictionaries (Kenny, 2018). It was not until over a decade later that the first ideas for using computers in translation were suggested by Andrew Booth and Warren Weaver, while the work on mechanical dictionaries continued to be developed through the 1940's and interest on mechanical translation continued to rise (Hutchins, 2003).

Throughout the first half of the 1950's this rise in interest persisted, with the publication of literature and project demonstrations on the subject raising awareness and allowing the establishment of ideas that are still relevant today, such as the need for post-editing on MT output, and the presentation of new ideas and future plans for this field of investigation (Hutchins, 2003).

In the second half of the 1950's, research on MT was separated between those who aimed at developing systems for mechanical translation and those who wanted to create a system that, in the long run, could produce satisfactory translations (Hutchins, 2003). It was also then that research approaches to MT began to be divided into three different methods: the Direct Translation model, the Interlingua model and the Transfer approach. The Direct Translation model consisted of creating specific programming rules for translation of one source language into one target language, while the Interlingua model required two stages of translation, in which the first would be to convert the source text into an universal interlingua, which would be the basis for all languages, and the second to convert the interlingua output into the target text. As for the Transfer approach, it consisted of converting abstract representations of the source language texts to their equivalent representations in the target language (Hutchins, 2003).

By the mid 1960's, MT research had expression in many countries all around the world. However, the optimism surrounding the development of MT gradually started to turn into disappointment, as the difficulty of the challenges that had to be surpassed in order to provide good quality MT output started to appear impossible to overcome (Hutchins, 2003). This wave of discontentment was further reinforced after the publication of a report requested by MT sponsors to assess the state of MT research. The ALPAC (Automatic Language Processing Advisory Committee) Report, published in 1966, dealt a heavy blow to investigation on the MT field, because it discouraged further investment by declaring there was no prediction of it ever being of use (Hutchins, 2003). The impact of the ALPAC Report was evident in the slowing down of MT research, which was an effect that lasted for a decade. However this does not mean that the work on MT stopped altogether during this period, with projects like TAUM (*Traduction Automatique de l'Université de Montréal*) still achieving important developments (Hutchins, 2003).

In 1976, research on MT started coming back to life and was mostly conducted around the Transfer-based approach. In the same year, the Commission of the European Communities purchased a version of Systran, a system developed by Peter Toma, and applied it to almost all languages of the European Communities. Since then, Systran has been applied at a group of intergovernmental institutions and companies⁷ (Hutchins, 2003). After Systran, other systems

⁷ Systran was originally a Russian-English system founded in 1968 to be used by the US Air Force. The system bought by the Commission of the European Communities was a French-English version and, later, the same system was developed for more European languages (Koehn, 2020).

started to appear which were meant for general usage but incorporated dictionaries for specific domains. The decades of 1970 and 1980 also saw the development of special-purpose systems and in the 1980's many MT research projects were founded all over the world (Hutchins, 2003).

However, satisfaction with the quality of MT systems' outputs was still a problem that had not been solved, and translators preferred to be set up with computer-aids for translation where they could be in control. Accordingly, through the 1990's many computer-based tools⁸ were developed in order to help translators (Hutchins, 2003).

At the same time, since 1989 a new approach to automatic translation had started to appear and resulted in what are called Corpus-based methods, that include methods such as the Example-based approach and the Statistical method (Hutchins, 2003). The Statistical Machine Translation (SMT) method is based on the idea that, rather than being based on linguistic principles alone, automatic translation can be produced through probabilistic models that are taught with bilingual corpora. Although this idea was not immediately accepted upon its proposal, it soon became the dominant frame for MT research in the 1990's (Kenny, 2018). The decade of the 1990's was also marked by the revival of investment in MT after the downfall caused by the ALPAC Report. At the same time, it also became standard for different research groups to share parallel corpora and make their software open source, which allowed for research and development to be much more open and accessible. This new mentality combined with the development of technology provided further impulse for the development of SMT in the 2000's (Koehn, 2020).

After 2015, SMT systems started losing dominance to a new approach to MT that outperformed previous systems - Neural Machine Translation (NMT). NMT systems work with artificial neural networks that are based on the human neural system. In other words, they consist of connections between sets of artificial neurons that depend on information passed between each other in order to be activated (Kenny, 2018). The training of neural systems for translation demands the attribution of weights to the artificial neurons that will allow them to be activated in layers (Koehn, 2020). NMT became the state-of-the-art in machine translation in recent years because it has allowed for a jump in the fluency of MT output. However, as pointed out by

⁸ As pointed out by Hutchins (2003), these consisted mostly of tools for concordancing, dictionary creation, terminology management and document translation and, later, of translator's workstations which combined software for all of these tasks and word processing.

Castilho (2019), this method is not without its flaws and, as such, its growth has been followed by a growing concern with improving and adapting strategies for the assessment of translation quality. While NMT systems produce extremely fluent outputs and perform well in relation to inflection and reordering (Toral & Sánchez-Cartagena, 2017), they are also more prone to hallucinations that result in target text that, while still fluent, can be completely detached from the source (Raunak et al., 2021), as well as omission errors. In addition, as NMT systems are dependent on training data, it is important to ensure that this data is of good quality. While it is important to have as much data as possible, it is also necessary that data quality is not poor, as this would result in similarly poor quality translations.

3.2. Quality Processes

The quality of machine translated output has been a concern from the beginning and one of the most defining factors when deciding if a certain system is acceptable for use. As such, throughout the years there has been a persistent effort to create translation quality assessment (TQA) processes that are adequate and efficient for MT tasks. The quality of translation can currently be assessed according to two main methods: manual quality metrics and automatic quality metrics, which will be explained more in detail in the following sections.

3.2.1. Manual Quality Metrics

Manual quality metrics to assess translation quality consist of methodologies in which the evaluator identifies errors in a translation according to specific error types, classifying them in order to attribute them a score. In the beginning, even though human evaluation of translation quality was carried out by experts, there was no pre-definition of standards for assessment in place, which meant that evaluation was subjective and, consequently, inconsistent (Lommel, 2018). One of the first attempts to standardize quality evaluation was the creation of score-cards in spreadsheets for evaluators to count errors and even scores. However, this method had two problems that made it inefficient: one was the fact that categorisation was still not standardized

across users and the other was that, because the errors were merely accounted for on a spreadsheet, it was difficult to link them to the translation (Lommel, 2018).

The appearance of the LISA QA Model in the 1990's and of the SAE J2450⁹ in 2001 marked the beginning of the creation of lists of error types to evaluate translations and, thus, standardization. The LISA QA Model defined a list of errors to be applied and a range of three severity levels (minor, major and critical) that were also to be attributed to them. In the end, the score of each segment was to be calculated according to the number of mistakes it contained and an overall score would also be attributed to the translation as a whole, additionally receiving a pass or fail evaluation according to it (Lommel, 2018). The LISA QA Model was a starting point for the customisation of many company-specific metrics, but soon it became evident that this system had limitations due to the fact that it took a one-size-fits-all approach, which made it difficult to adapt to specific tasks (Lommel et al., 2014).

In 2012, after LISA's dissolution in 2011, the Translation Automation User Society (TAUS) started developing DQF (Dynamic Quality Framework) in an attempt to shift the resolution of problems in translation to before the beginning of the process, including the definition of quality requirements (Castilho et al., 2018). In order to create their own error typology, TAUS gathered information from language service providers (LSPs) to release a typology that was relevant to their needs. Upon its release, the DQF Error Typology had six error types, divided into further subtypes, four categories to identify other issues and four different severity levels that are similar to MQM (Lommel, 2018).

In 2015, the EU-funded research project QTLaunchPad also began working on TQA, namely creating the MQM framework that has been mentioned before in **Section 2.2.1**. While MQM was developed partly in order to address the limitations of the LISA QA Model, at the same time it is built on many of its principles (Castilho et al., 2018). MQM defines a set of error types that can be adapted by the users, according to what is relevant to them, allowing them to create their own MQM-based error typologies. MQM is a hierarchic framework with four maximum layers in which every issue can be used as an error tag. In total, MQM has a total of over 100 issue types at different levels of granularity, with each main issue type containing several daughter issue types that are further divided into granddaughter issue types. In addition,

⁹ The SAE J2450 consisted of a simple metric based on the score-card method that included a total of six error types and two severity levels.

MQM includes the severity levels of *Minor*, *Major*, *Critical* and *Null*. With the exception of the *Null* severity, all the other levels correspond to different weights. *Minor* errors have a weight of 1, while *Major* errors weigh 5 points and *Critical* errors correspond to a weight of 10 points. These points are deducted from a total score, resulting in an MQM score that is obtained with the following formula:

$$=100-\text{SUM}((1*\text{MINORS})+(5*\text{MAJORS})+(10*\text{CRITICALS}))/\#\text{Words}*100$$

This system allows users that create MQM-based error typologies to decide what level of granularity they want to allow when tagging errors. MQM creators recommend that metrics should avoid being too fine-grained, as this can make different categories difficult to set apart, creating ambiguity in the process of annotation and, as a consequence, affecting IAA¹⁰. In light of this there is also MQM Core, another version of MQM which is more simplified, containing only two levels of issue types and 7 main categories, which amounts to a total of under 30 issue types¹¹ and a new version of MQM¹² was also released in 2022.

By 2014 the TAUS DQF and the MQM framework were being developed separately, which was not efficient from the point of view of fast development. As such, in 2015 the two frameworks were integrated with each other as per QT21, QTLaunchPad's follow-up project. The integration of these frameworks demanded changes in format from both sides and resulted in the Harmonized DQF-MQM framework, which has less issue types than the MQM hierarchy but still allows customization. These characteristics make the DQF/MQM framework more simple to apply and to understand and, as such, allowed it to become the most popular method of applying MQM in recent years (Lommel, 2018).

Aside from MQM, Direct Assessment (DA) is also widely used by researchers in the MT field. As proposed by Graham et al. (2013) DA consists of using human judgment to determine the quality of MT outputs on a scale from 1 to 100 in relation to a reference translation. However, due to the fact that it is a type of evaluation that is not performed based on a

¹⁰ <https://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html>

¹¹ <https://themqm.info/typology/>

¹² <https://themqm.org/>

standardized error typology, the resulting data is not very fine-grained in the sense that it does not reflect the specific errors of the MT output.

3.2.2. Automatic Quality Metrics

Although manual quality metrics can often provide results that are more fine-grained, they also consume large amounts of time and human resources, in addition to being a very expensive process. This means that LSPs responsible for large volumes of translations are unable to evaluate all translations using manual metrics. Similarly, it is impractical in the process of continuous improvement of systems which requires frequent and fast evaluations. As such, the adoption of automatic quality metrics has become widespread, both on their own and combined with manual metrics.

The first automatic metric to be created was BLEU in 2002, which was developed by the IBM group (Hutchins, 2003). BLEU stands for Bilingual Evaluation Understudy and it works on a reference-based method, meaning that it calculates a quality score based on how the MT output correlates with reference translations previously produced by humans (Papineni et al., 2002). However, BLEU's reliability is sometimes questionable. This is due to the fact that high evaluation scores are heavily dependent on the match of word sequences between source and target, penalizing paraphrasing instances regardless of whether they are incorrect translations. (Castilho et al., 2018).

One of the metrics to be developed after BLEU was METEOR (Metric for Evaluation of Translation with Explicit ORdering) in 2005. METEOR is “an automatic metric for machine translation evaluation that is based on a generalized concept of unigram matching between the machine produced translation and human-produced reference translations” (Banerjee & Lavie, 2005:1). METEOR also addresses some of the problems of BLEU by allowing the use of synonyms and paraphrasing instead of rigid adherence to references (Castilho et al., 2018). The score is obtained based on how the unigrams match between source and target and, like BLEU, the evaluation of the metric is done by analyzing how that score correlates to human judgment (Banerjee & Lavie, 2005).

Following the development of Neural Machine Translation systems and the consequent improvement in machine translation quality, metrics like BLEU started becoming insufficient to assess the quality of those translations. This is due to the fact that NMT outputs are often not as straightforward, which for evaluation metrics means the correlation between source and target may be low and produce low scores. As such, a new neural framework for MT evaluation called COMET (Crosslingual Optimized Metric for Evaluation of Translation) was released by Unbabel in 2020 in order to address the fact that “current metrics struggle to accurately correlate with human judgment at segment level and fail to adequately differentiate the highest performing MT systems” (Rei et al., 2020:1). COMET is “a neural framework for training multilingual machine translation evaluation models” (Rei et al., 2020:1). Differently from previous metrics, instead of using only reference translations as a basis for assessment of quality, COMET also integrates the source input into the evaluation models (Rei et al., 2020). This has resulted in COMET obtaining new high levels of correlation with human judgment. COMET is trained with three different types of data, which results in three versions of this system: COMET-HTER, which is trained with the QT21 corpus, COMET-MQM, trained with an internal MQM corpus, and COMET-RANK, trained with the WMT DARR corpus from 2017/2018 (Rei et al., 2020). When compared to other baseline metrics such as BLEU, all three different models of COMET were successful in obtaining new state-of-the-art results in terms of correlation with human judgment (Rei et al., 2020), establishing COMET as the new state-of-the-art in MT evaluation as well as becoming the main automatic quality metric used at Unbabel.

3.3. East Asian Languages in Machine Translation

Due to the fact that the objective of the work on this thesis was to build an annotation module that can ultimately improve the manual quality assessment processes for Japanese, Korean and Chinese within Unbabel, it is important to first understand the work that has been done so far concerning these languages in the scope of machine translation.

As mentioned in **Section 3.1.**, before the publication of the ALPAC Report (1966), which slowed down the MT investigation movement in 1966, research on machine translation had already spread across the world, including countries such as China and Japan (Hutchins, 2003).

In fact, after the effect of the ALPAC Report (1966) started to subside both countries continued to have a visible role in the MT research field.

In 1972 the Chinese University of Hong Kong started to develop the CULT system (Hutchins, 2003). This system was designed for the purpose of translating mathematics texts from Chinese into English and it started to be used regularly for this end in January of 1975 (Loh & Kong, 1977). Since 1972 and in years following its implementation, this system continued to be improved to allow translations in the English to Chinese direction as well and was adopted as a basis for an online MT system which was put in use at the Chinese University of Hong Kong (Loh & Kong, 1979).

As explained by Hutchins (2003), in the decade of 1980 Japan dominated the computer-aided translation field in the sense that many national computer companies at the time were invested in producing software for this purpose. These systems were usually built for specific fields of translation and required much pre- and post- editing efforts and were, naturally, focused mostly on the English to Japanese, and vice-versa, translation directions (Hutchins, 2003). In addition, as Nagao (1989) points out, during the 1980's Japan was prolific in the field of machine translation, developing other projects such as designing MT systems for languages of surrounding countries and working on electronic dictionaries, projects which were carried out not only by software companies but also in universities which also participated in this field of research. In fact, much of this activity surrounding MT research in Japan in the 1980s started under the influence of the Mu project which was developed at Kyoto University (Hutchins, 2003). It is also important to mention that since this decade Japan concentrated efforts into developing example-based translation systems before the rise of SMT (Koehn, 2020).

The 1980s also saw the founding of more research projects in Korea and mainland China (Hutchins, 2003). This time marked the beginning of MT research in Korea, starting in universities under the influence of the rise in interest around artificial intelligence (AI), which also prompted the Korean government to invest in the development of machine translation later in the decade (Park & Oh, 1999) as part of the wave of revival in interest in the field mentioned by Hutchins (2003). This resulted in the development of several systems which were mostly concentrated in the translation directions of Japanese to Korean and English to Korean, with this second language pair being developed slightly later (Park & Oh, 1999). However, MT research

in Korea slowed down again towards the end of the 1990s due to the fact that the quality of the translations produced by these systems was below what users expected (Park & Oh, 1999).

As SMT became the focus of MT research in the 1990's, the need for monolingual and bilingual corpora also grew and continues to be of importance in the age of NMT. However, as pointed out by Tian et al. (2014), as late as 2014 the development of publicly available English to Chinese parallel corpora was still not on par with that of other language pairs. The same can be said about Korean and Japanese, as Japanese-English corpora were also reported to be in shortage in the decade of the 2000's (Ishisaka et al., 2009) and much of the currently available open corpora for Korean was developed only after 2018 (Cho et al., 2020). However, it should be noted that there is a crescent effort to build publicly available NLP tools dedicated to these languages, including parallel corpora (Diño, 2019).

In addition to the efforts concentrated on the improvement of NLP tools for East Asian languages, it is also important to consider the recent MT systems developed specifically for translation of these languages. In the case of Chinese, the NMT systems developed by Sogou (Wang et al., 2017), Baidu (Sun et al., 2019) and Tencent (Wu et al., 2020) which were presented for the translation tasks at WMT17, WMT19 and WMT20, respectively, should be highlighted. For Japanese and Korean there are also important developments such as the NMT systems proposed by Rakuten (Susanto et al., 2021) and the combined system of SMT and NMT used by Naver (Lee et al., 2015), respectively.

Finally, from the point of view of quality processes, recent years also saw the development of various studies, particularly regarding manual quality metrics, as is the case of the annotation typology for the translation direction of English to Chinese proposed by Yuying Ye and Antonio Toral (2020), which will be described in **Section 3.4.2.** and was an important basis of comparison for the work developed on this thesis.

3.4. Error Typologies for Annotation

As will be discussed in further detail on **Section 4** of this work, the main objective of the project developed during this internship was the creation of a module for annotation of translation errors suited for East Asian languages. After the creation of the module, it was

important to test its effectiveness and it was decided that the best approach was to annotate the same datasets using two different typologies, in addition to the one that is proposed in this thesis. The purpose was to analyze how different typologies have an impact on the quality of annotations and to investigate the benefits of having a dedicated typology, in this case for a specific subset of languages, through concrete data. Although the results of this investigation will only be discussed in **Section 5**, it is important to firstly present the typologies that will be used in this comparison.

The taxonomies that will be used to carry out this investigation are the error typology that was in use at Unbabel until June of 2022, which will be presented in **Section 3.4.1.** and the error typology proposed by Yuying Ye and Antonio Toral (2020), which was chosen for this comparison due to the fact that it was developed specifically for annotation of translations in the English to Chinese direction and, as such, constitutes an effort at creating an annotation typology dedicated to an East Asian language. This typology will be described in **Section 3.4.2.**, along with some of the most important findings relating to it that were later relevant while building the annotation module proposed in this thesis.

3.4.1. Unbabel's Error Typology

The first typology that was used in this investigation was the MQM-compliant annotation typology that was in use at Unbabel until June of 2022 (v2), when it was replaced by a new version (v3) created due to the integration with Lingo24 previously mentioned in **Section 2.3.** Since this typology was commonly used for annotation tasks corresponding to all language pairs offered by Unbabel, the annotated data that was analyzed leading up to the creation of the annotation module was all related to the former and it made sense that the datasets for this experience were annotated using it as well.

This error typology, as represented in **Figure 3**, contains 3 coarse categories and 16 parent issue types, 30 daughter issue types and 10 granddaughter tags, amounting to a total of 47 error types that can be selected, as parent issues are not selectable.

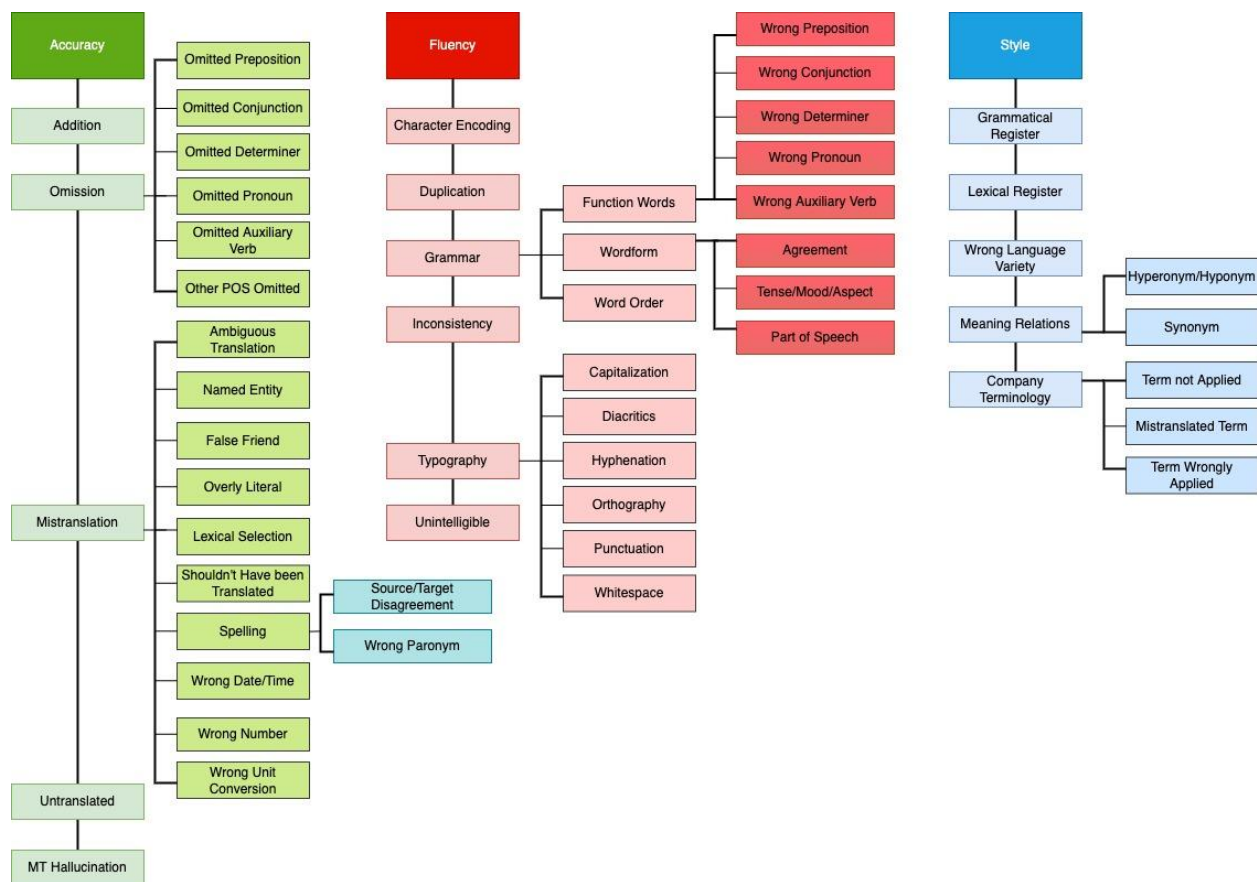


Figure 3. Unbabel Error Typology

3.4.2. Asian Languages Focused Typology

The second typology that was tested was the MQM-compliant error taxonomy proposed by Yuying Ye and Antonio Toral (2020), which was designed to be used in the translation direction of English to Chinese and, as such, intends to be adapted to its relevant issues. As stated by the authors, this taxonomy was created to provide “a detailed human analysis of the outputs produced by state-of-the-art recurrent and Transformer NMT systems” and the analysis was carried out in the news domain.

The full taxonomy is represented in **Figure 4** and it contains a total of 15 selectable error types.

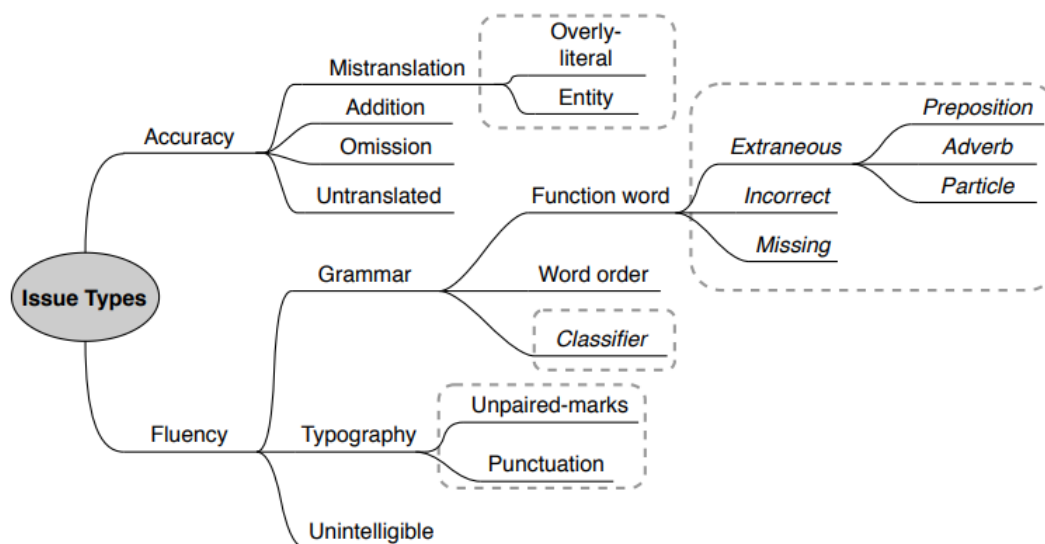


Figure 4. The MQM-compliant error taxonomy for the translation direction English-Chinese (Ye and Toral, 2020)

It is important at this stage to point out the biggest differences between this typology and the one previously discussed in **Section 3.4.1.**, in order to better understand its general limitations and those that surface when applied to translations in the Customer Support domain. This will serve as the basis for interpreting and discussing the results in **Section 5**, where the annotations for the three typologies will be compared and analyzed.

The first difference that is important to point out is the extension of the typology itself. The typology proposed by Ye and Toral only has the two coarse categories of *Accuracy* and *Fluency*, and while the number of parent tags is not much different from the Unbabel Error Typology, the fact that the latter is considerably more fine-grained means that Ye and Toral’s typology has less issue types to choose from. At the same time, the typology in use at Unbabel has an entire category that is not included in the typology proposed by Ye and Toral (*Style*) and none of the issue types under it are included in Ye and Toral’s typology, which means that from the start it was accepted that this typology would at least not allow the annotation of *Whitespace* and *Register* errors. It also does not contain anything related to *Terminology*, but this could potentially be annotated using other tags, as will be discussed later in **Section 5**.

Finally, another difference which is obvious due to the fact that the typology described in the present section is aimed at the translation direction of English to Chinese, is the fact that there are both issue types that were intentionally not included and others that are present due to the fact that they are relevant in the context of this translation direction and that would not make sense in a typology developed from an exclusively western point of view. For example, error types such as *Spelling* were not included in this typology due to its lack of relevance in the context of the above mentioned translation direction. However, this typology also introduces error types related to important components in Chinese language, such as particles and classifiers, which were extremely relevant in creating the annotation module proposed in this work and will, therefore, be discussed more in depth in **Section 4.3.1.1.**

This chapter presented briefly the history of the development of machine translation and the processes used for assessing translation quality. In addition, the development of the same was also discussed from the point of view of the East Asian languages under discussion in this thesis. Finally, this chapter presented an overview of the two typologies which will be used to compare with the annotation module proposed in this work.

4. Methodology

In the context of a company that works with multiple languages with completely different structures, as is the case of Unbabel, it is extremely difficult to create an one-size-fits-all annotation system. At the same time, it is also hard to use methodologies such as MQM due to the fact that they are biased towards Western languages and their grammatical structure. As such, it is challenging to conceive a general methodology that can be equally applied to the languages under discussion in this thesis.

At the same time, as mentioned in **Section 3.2.1.**, it is recommended that MQM based typologies should not be overly fine-grained. This is because human annotation is always going to be susceptible to some degree of subjectivity so, in order to avoid any ambiguity that might affect IAA, it is necessary to not overwhelm the annotators with an excessive amount of options if the distinction between certain categories is not essential. In face of this, it can be concluded that the particularities of each language cannot, and should not, be addressed within one single typology.

Nonetheless, it remains true that some of the details that are left behind when building a general typology can have a negative impact on the quality of the annotations and on the consistency between annotators of specific languages, making it difficult to use the resulting data reliably. Such is the case for the East Asian languages¹³ Unbabel works with, which have had recurring annotation problems partly due to the fact that the current typology lacks some details that can have a considerable impact on the quality of the annotations for these languages, concerning both the error tags and the guidelines.

In light of the above, **Section 4.1.** will further discuss the main objective of the work developed during this internship, which was the creation of an annotation module that was specifically adapted for the annotation of the East Asian languages previously mentioned.

¹³ For the purposes of this thesis, the East Asian languages that are referred to will be Japanese, Korean and two varieties of written Chinese - Simplified and Traditional - which will be henceforth treated as two separate languages due to the fact that they are also handled as so within the company and that they presented different issues in terms of the annotation process.

Following that, in **Section 4.2.** the tasks executed at Unbabel during this internship will be described. Finally, **Section 4.3.** will present the East Asian Languages Annotation Module for the Unbabel Quality Framework this thesis proposes and the process of its conception, by first exploring the annotation problems that were found through analysis of real Unbabel data and that motivated this project in **Section 4.3.1.**, and then presenting the finished annotation module and its respective guidelines in **Section 4.3.2.**

4.1. Objectives

The main objective of the work developed during this internship was to design and test an annotation module that considers the specific characteristics of the East Asian languages that constitute the object of work on this thesis, in an effort to improve the quality and consistency of the annotations for these languages, which have thus far suffered the impact of being performed with typologies that are not entirely adequate.

The annotation module proposed in this work aims to solve the shortcomings that the MQM framework (Lommel et al., 2014) and, more specifically, the MQM compliant typology in use at Unbabel present in the context of annotating these languages, as they were developed from an European perspective and do not cover some important specificities of East Asian languages, resulting in annotations that do not reflect accurately the quality of a translation.

Although in recent years there have been efforts to address this issue, such as the typology proposed by Ye and Toral (2020) that was presented in **Section 3.4.2.**, the error taxonomy proposed in this thesis aims at being more detailed, in order to cover more issue types that are necessary for annotating Customer Support content, as well as other errors that were considered important from a general point of view, while still maintaining that an error typology should not be overly fine-grained to avoid confusing annotators and allowing too much subjectivity.

4.2. Internship Tasks

The tasks performed during this internship were focused on exploring the existing issues around East Asian languages, in order to not only determine what problems the annotation module for these languages should address, but also to help improve the information about these languages available within the company and to the community of freelancers. In addition, as will be discussed in more detail in **Section 4.2.1.**, support was also provided during the process of the integration with Lingo24.

Firstly, it was important to get familiar with both the quality processes at Unbabel which were already introduced in **Section 2.2.** and the scope of work of the teams this internship project is inserted in. At the same time, since the main objective of the internship was to create an annotation module, it was essential to get fully acquainted with Unbabel's Annotation Tool, the general annotation guidelines available to every annotator and the specific language guidelines for each of the languages the annotation module is directed at.

After this process of familiarization, there were enough basis to analyze real annotation data in order to assess which type of issues were affecting the quality of the annotations. This was carried out in three main steps:

- 1) conducting a preliminary analysis of already existing annotated data;
- 2) evaluating annotations and providing individual feedback to annotators;
- 3) first-hand annotation using the four languages under analysis.

After collecting data on which issue types the annotators were struggling with or using incorrectly in the first two steps, annotating the same type of content first-hand served the purpose of further determining which type of errors proved difficult or impossible to annotate correctly, both in relation to the typology itself and the annotation guidelines. Only once this data was collected and properly analyzed it was possible to start designing the annotation module which aims at solving the problems that were found.

In addition to the work mentioned above, which was directly related to the creation of the annotation module, support was also provided in other areas within the company concerning the languages under discussion in this thesis. These tasks included discussing causes and potential solutions for poor quality MT outputs concerning these languages, particularly Japanese, and

providing insight on language specific issues, such as whitespace rules and tokenization challenges.

4.2.1. Preprocessing of Asian Languages

In the context of NLP, tokenization consists of the process of separating sentences into units and attributing them to specific functions before a translation progresses to the stage of annotation. The opposite process, de-tokenization, consists of returning the units to their original state.

During the duration of this internship help was provided in improving the de-tokenizers for East Asian Languages at Unbabel, particularly in the case of Korean. While the Korean language uses whitespaces, as opposed to Japanese and Chinese, the whitespace rules are complex as different units do not necessarily have to be separated. This poses significant challenges in terms of the tokenization process as it means that not only words are often separated incorrectly but, because of this, they are attributed the wrong function. As such, during this internship insight was provided in relation to whitespace rules in Korean and identification of the correct Part of Speech when needed in order to improve the performance of the detokenizer.

4.2.2. Quality Annotation for Lingo 24 Integration into Unbabel

The integration of Lingo24 with Unbabel meant that the company greatly expanded the number of language pairs offered and that customer support content was no longer the exclusive object of translation Unbabel is responsible for. The process of integration took place during this internship and, as such, help with the process was provided when necessary, namely relating to issues with the languages mentioned in this thesis.

Due to the fact that the content introduced by Lingo24 is different and more focused on the marketing aspect, one important task was to assess how prepared the annotators from Unbabel were to annotate Lingo24 content. At the same time, help was also provided in

comparing Unbabel and Lingo24 MT engines, which was important in the process of reaching quality parity between the two.

Lastly, it was important to conciliate the existing language guidelines at Unbabel with specific guidelines from Lingo24 clients which, as mentioned above, often require more technical translations and other complex localization challenges.

4.3. Methodology for Error Annotation for East Asian Languages

A preliminary analysis of existing annotation data for the languages discussed on this thesis revealed that, although the expected quality of MT for chat is relatively low, the MQM scores were fairly high while not corresponding to actual high quality translations. Due to the fact that MQM scores depend on annotation, this meant that there were issues in the annotation process for these languages.

As such, further investigation was conducted by analyzing several datasets of all four languages with the objective of determining which aspects of the annotation process were having a negative impact on the reliability of these annotations. This investigation was conducted in two main parts that will be discussed in detail in **Section 4.3.1.**, and which, in summary, aimed to determine the problems that were related to the annotation typology in use at Unbabel, which was explained in detail in **Section 3.4.1**, and the problems related to the usability of the annotation software itself.

This investigation revealed that most of the issues found within the annotation data related to the fact that the annotation typology and corresponding guidelines were not entirely adequate for annotating these languages. Aside from inconsistent and inappropriate attribution of severities to each error and under-annotation issues, which all contribute to high MQM scores that do not reflect the true quality of a translation, it was found that a large amount of the mistakes could be attributed to the lack of appropriate error types to annotate them and to the lack of some detail regarding important particularities of these languages in the annotation guidelines. In light of this, it became clear that it would be beneficial to have an annotation typology and respective guidelines that would be built while keeping in mind the specificities of these languages. As such, **Section 4.3.2.** will explain the process of building the East Asian

Languages Annotation Module for the Unbabel Quality Framework which aims at being suited for annotation of Japanese, Korean, Simplified and Traditional Chinese and focused on a data-driven approach..

4.3.1. Annotation Challenges

As mentioned before, in this section the quality of annotations will be analyzed in two phases. The objective of the first phase was to obtain a first assessment over what types of problems existed in the annotation for the languages the work on this thesis relates to and to determine their origin. The dataset used for this analysis consisted of 342 segments from previously annotated ticket, or e-mail, data performed by one annotator for the translation direction of English to Simplified Chinese while using the Unbabel Error Typology, which was introduced in **Section 3.4.1.** **Figure 5** illustrates the results of this analysis by showing the percentage of each type of mistake the annotator made in relation to the total number of mistakes.

Annotation Errors

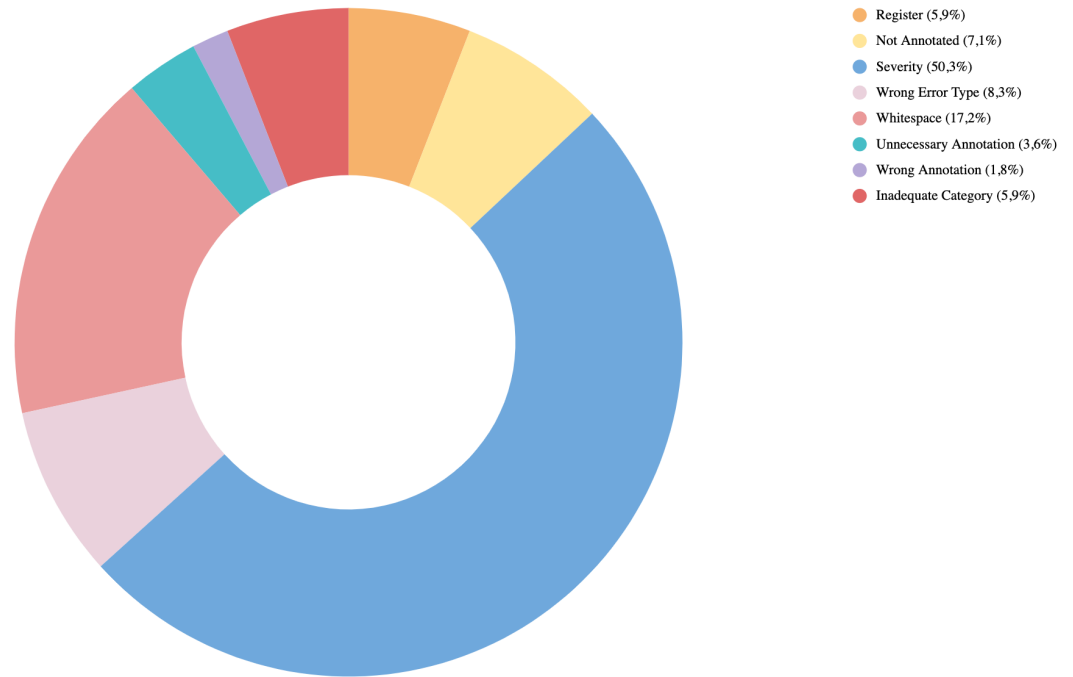


Figure 5. Percentage of annotation errors in Simplified Chinese dataset

Through this analysis it was possible to determine which types of annotation mistakes were most prolific as a first step to understand what needed to be addressed in the annotation module. However, before explaining the results of this investigation in more detail it is necessary to firstly specify what types of mistakes were evaluated and how they were defined:

- Any annotation that was identified with the wrong issue type was marked as **Wrong Error Type**, unless the error is derived from the lack of an appropriate issue type to annotate it in the typology, in which case it is identified as an **Inadequate Category** mistake.
 - Wrong Error Type

EN (source)

(1a) Hello John_Smith,

ZH-CN (target)

(1b) 您好 **John _ Smith**,¹⁴

✗ [Punctuation]

✓ [Named Entity]

According to the guidelines for the Unbabel Error Typology, which was the typology used for these annotations, any error that falls upon a Named Entity should be annotated as *Named Entity*, even when the error could seemingly be annotated using other issue types, such as *Punctuation*, *Whitespace* or *Capitalization*.

- Inadequate Category

EN (source)

(2a) An email will be sent for the survey purposes.

ZH-CN (target)

(2b) 出于调查目的，我们将向您发送调查电子邮件。

[Omitted Determiner]

→ Proposed issue type in the annotation module: *Omitted Classifier*

As the annotator points out in the comments regarding this segment, what is omitted is a measure word, or classifier. However, the Unbabel Error Typology does not cover such errors specifically, instead indicating the annotators should use the *Omitted Determiner* issue type instead. As such, this is not a mistake made by the annotator, as they are complying with the guidelines, but rather an error that originates from the lack of an appropriate issue type, which makes the annotation grammatically incorrect and can lead to confusion due to not being appropriate for these types of errors. These errors will be discussed more in depth in **Section**

¹⁴ The examples in this thesis will be, in their majority, highlighted in bold to indicate what portion of the segment was annotated and color coded to demonstrate the severity that was attributed to the error. Yellow corresponds to minor errors, orange to major errors and red to critical errors.

4.3.1.1., where the new categories added to the East Asian Languages Annotation Module for the Unbabel Quality Framework will be discussed.

- **Wrong Annotations** are those that are incorrect in the sense that they should not exist (for example annotation of an addition error in the target for a part-of-speech that exists in the source and its removal in the target would affect the meaning), while **Unnecessary Annotations** are those that are not necessarily wrong but are not needed because the target text is already correct.

- Wrong Annotation

EN (source)

(3a) I kindly request you to try the below troubleshooting steps and help me resolve the issue.

ZH-CN (target)

(3b) 请您尝试以下故障排除步骤，并帮助我解决这个问题。

[Addition]

This is the case of a wrong annotation, in the sense that it should not have been made because it would mean altering the meaning of the source. In this segment the annotator suggests the removal of the “help me” part in the target, which would change the meaning of the sentence and, as such, is a wrong annotation.

- Unnecessary Annotation

EN (source)

(4a) We see that you are concerned about a few items you were supposed to receive as part of the events/promotions.

ZH-CN (target)

(4b) 我们看到您对一些您应该收到的物品感到担忧, 作为活动/促销的一部分。

[Addition]

In this case the comma was annotated as an error. In addition to constituting a Wrong Error Type mistake, as the issue type used should have been *Punctuation*, this annotation is not necessary. Removing the comma in this segment would not compromise the meaning of the sentence, but its existence is not wrong therefore it should not have been tagged. It is important to note that over-annotation creates unnecessary noise in annotated data, so annotators are advised to avoid doing this and to only annotate what is objectively wrong, rather than annotating based on their preferences.

- Errors in the target that the annotator missed were marked as **Not Annotated**¹⁵.

EN (source)

(5a) Hello MARY,

ZH-CN (target)

(5b) 您好 MMARY ,

While this segment was annotated with the *Word Order* and *Punctuation* issue types, there should also be a *Named Entity* annotation, as the named entity in the target is spelled incorrectly. The *Named Entity* annotation would be the most important one on this segment and would also carry the heaviest severity due to being the most impactful error. Instances like this have to be noted due to the fact that their non-annotation influences MQM scores to be higher while not corresponding to the true quality of the translation at hand.

¹⁵ This dataset contained many punctuation errors related to the fact that Chinese punctuation is meant to be full-width and in the target text it is mostly written in half-width. As the annotator did not mark any of these errors, the situation was taken note of and addressed in the guidelines for the new proposed annotation module but due to the fact that there was a very large amount of them, this issue was not taken into account for the purpose of this analysis, as it would be a very heavy percentage in the Not Annotated category.

- **Severity** mistakes are those that occur when the wrong severity is attributed to a certain issue type in a specific context.

EN (source)

(6a) Hello [NAME],

ZH-CN (target)

(6b) [NAME] , 你好!

[Grammatical Register]

In **example 6**, the target text was translated informally while the job required translations to be in a formal register. The annotator attributed a minor severity to this error which should have been major, due to the fact that translations in registers lower than they are supposed to be may be considered offensive.

- Finally, it is important to point out that **Whitespace** and **Register**¹⁶ are actual issue types in the typology and that their misuse could be marked as Wrong Annotation. However, since it was evident from the start that as individual mistakes they occurred more than others, they were analyzed separately.

- Whitespace

EN (source)

(7a) 1. tap the icon in question

ZH-CN (target)

(7b) 1 .点击相关的图标

¹⁶ The version of the Unbabel Error Typology that was used for annotating this dataset (v2) distinguishes between Lexical Register and Grammatical Register but for the purpose of this analysis both categories were considered together.

This segment contains a *Whitespace* error right after the numeration in the target but the annotator has not marked it.

- Register

EN (source)

(8a) I have removed the flag and you should be able to make a payment now.

ZH-CN (target)

(8b) 我已经删除了标记，您现在应该能够进行 付款。

The register required for this job was informal, yet the segment was translated formally and a pronoun of the wrong formality was used, which is an error that should have been annotated.

As shown in **Figure 5**, half of the annotation mistakes in this dataset is related to the attribution of inadequate severity levels to errors. This could mean that the annotator had a poor understanding of the annotation guidelines or that the guidelines themselves were not specific enough regarding attribution of severities. This became more evident when further analysis of the annotations revealed that some of the exact same mistakes in the same context were being attributed different severity levels, as seen in **examples 9** and **10**.

Annotation 1:

EN (source)

(9a) 1. tap the icon in question

ZH-CN (target)

(9b) 1. 点击有问题的图标

[Lexical Selection, minor]

Annotation 2:

EN (source)

(10a) 1. tap the icon in question

ZH-CN (target)

(10b) 1. 点击有问题的图标

[Lexical Selection, major]

Following Severity, *Whitespace* was the most prolific type of mistake, specifically the non-annotation of *Whitespace* errors in the target. As mentioned before, due to the fact that there was a considerable amount of *Whitespace* errors in this annotation dataset they were considered separately, instead of being included under “Not Annotated”, in order to obtain a better overview of the most important types of mistakes.

The instances of non-annotation, register related mistakes, using the wrong error type and mistakes that can be attributed to the lack of an adequate typology all share a similar weight. In the case of *Register*, similarly to the approach used for *Whitespace*, the mistakes marked as such are related to the non-annotation of *Register* errors existing in the target.

Regarding the errors that lack proper error types to classify them, these situations are identified as “Inadequate Category” cases. For this dataset only the errors related to measure words¹⁷ were identified, although there are other types of errors that can be included in this category and that should be discussed. The annotator of this dataset chose to identify these errors as *Omitted Determiner*, which is correct according to the language guidelines Unbabel provides but is not the most correct way to annotate these words and can lead to confusion among annotators and subsequent inconsistency in annotations due to each annotator choosing the issue type they feel is more adequate.

In fact, inconsistency due to the lack of issue types that are commonly agreed to be the most adequate seems to be one of the biggest consequences of using a general typology for annotating some specific languages. This forces annotators to make decisions that are more subjective than desirable, producing ambiguous results and affecting IAA scores negatively since the annotation becomes dependent on personal opinion. Although the percentage of mistakes generated due to this factor is not very high comparatively to other types of mistakes, it was

¹⁷ The issue around measure words will be discussed more in depth in **Section 4.3.1.1.**

believed that annotation quality could greatly benefit from having this problem solved, which is one of the main motivations that drove the creation of an annotation module specific for East Asian languages. The issue types that were considered to be missing in the analyzed datasets and that were included in the annotation module will be discussed in more detail in **Section 4.3.1.1.**

Due to the fact that all the annotations analyzed in this dataset were made by the same annotator, this analysis did not necessarily apply to all the annotators for this language pairs, much less to the annotation problems of all the other languages the annotation module is meant to be applied to. As such, in the second phase of analysis it was important to conduct an investigation that encompassed all languages and more annotators to get a more detailed understanding of the problems that needed to be addressed.

Figure 6 illustrate the results of the analysis of tickets annotation datasets in all four language pairs: English to Japanese (1172 nuggets), English to Korean (1903 nuggets), English to Simplified Chinese (570 nuggets) and English to Traditional Chinese (573 nuggets).¹⁸

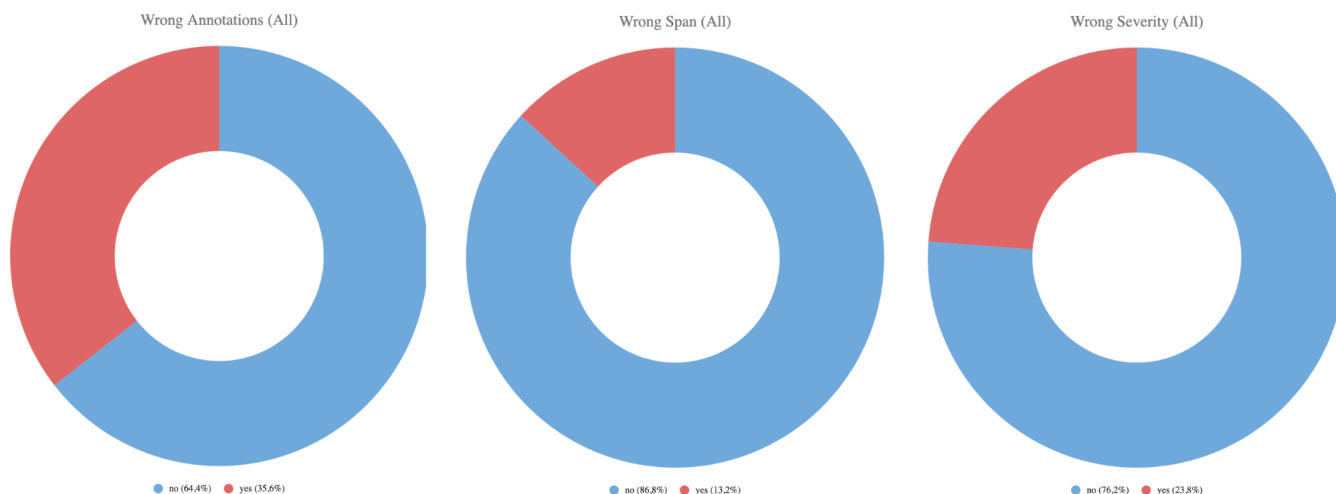


Figure 6. Percentage of annotation errors for all LPs

¹⁸ It was preferential to use annotation data from more than one annotator for each language pair but this was not always possible. As such, this dataset was annotated by 5 Japanese annotators, 3 Simplified Chinese annotators, 2 Traditional Chinese annotators and 1 Korean annotator.

The purpose of this analysis was to determine what types of annotation mistakes the annotators for these languages do more frequently, in order to pinpoint existing problems and attempt to solve them. As such, errors in annotation were separated into the three categories of wrong annotation, wrong span¹⁹ and wrong severity and each segment was marked with *yes* or *no* depending on whether it contained an annotation mistake belonging to any of those categories.

1	lp	source	target	typology_error	error	severity	wrong annotation	wrong span	wrong severity
20	en_ja	If you encounter any issue please send us a screenshot of the error and we will take a further look.	問題が発生した場合は、さらに調査させていただきますのでエラーのスクリーンショットをお送りください。	punctuation	でエ	minor	no	yes	no

Figure 7. Example of annotation analysis

Figure 7 shows how this analysis was conducted. Each segment was reviewed and classified according to whether the annotator made an annotation mistake (*yes*) or not (*no*) in each field. In the example shown, the annotator has made a correct annotation for a punctuation mark that is missing and attributed the right Severity to the error, so the corresponding fields were marked with *no*. However, they annotated this error using the incorrect span, as they selected the two characters in between which the punctuation mark should be instead of a full word as the guidelines instruct, therefore the wrong span category for this segment was marked with *yes*.

Following these categories, out of the approximately four thousand annotations in the dataset, around 29% of had some type of issue²⁰. **Table 1** shows the percentage of annotations that were wrong for each type, overall and across the different languages.

¹⁹ The Span of an error corresponds to its length. For annotation purposes, the minimum span that can be selected should be a whitespace or a whole word, while the maximum can be an entire segment.

²⁰ Errors that were annotated with register issue types were removed from the dataset because the information regarding the intended register was not available and could not be verified. As such, the data presented does not account for those issue types.

	Wrong Annotation	Wrong Span	Wrong Severity
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 1. Percentage of annotation errors per LP

From **Table 1**, we can see that there is a clear distribution of the most common types of errors between Japanese/Korean and the two variants of Chinese. Overall, the biggest percentage of mistakes made by the annotators corresponds to wrong annotations. This means that the wrong error type has been attributed or that an error was tagged when the segment did not need to be corrected. However, this percentage is heavily influenced by the data on Korean, since for this language almost 60% of the annotator’s mistakes are wrong annotations. As for Japanese, where wrong annotations are also the most common mistake, the percentage is much more balanced in relation to Wrong Span and Wrong Severity mistakes. This, however, means that there is an overall struggle with annotating Japanese.

In the Chinese annotations, both for the simplified and traditional datasets, the most common mistake is the attribution of the error severity to the errors, with the difference compared with the other types of mistakes being especially relevant in Traditional Chinese. It is important to note that correctly attributing appropriate severities to each error is important due to the fact these heavily influence MQM scores.

Overall this analysis revealed that the annotators were making more mistakes than desirable and it was important to investigate once more the possible causes behind this, in order to find the best way to solve the issues.

Since this analysis did not discriminate between types of mistakes within each category like the previous one, many of those types of mistakes are included under “Wrong Annotation” here and it becomes evident that this is the most common mistake across all datasets. As

mentioned before and by analyzing this dataset in detail, it is possible to affirm that many of these mistakes are related to two major factors: the lack of appropriate categories and confusion in relation to the annotation guidelines. This is due to the fact that these East Asian languages have components that were not accounted for in the Unbabel Error Typology used to annotate these datasets, such as particles and measure words, which will be discussed in further detail in **Section 4.3.1.1.** This makes it so that annotators often have to decide on their own which category to apply, resulting in mistakes and poor agreement. At the same time, the language guidelines are not clear in some specific points that were taken note of and addressed in the annotation guidelines for the annotation module presented in **Section 4.3.2.1.**, which also contributed to the lack of quality of these annotations.

However, it should be noted that while wrong annotations are a problem that exists across all languages in this dataset, the overall percentage is heavily affected by the fact that almost 60% of the Korean annotations suffer from this problem. The specific problem with Korean in this case comes from a combination of two factors:

- As pointed out before, the Korean dataset is the largest, but it also has only one annotator.
- The annotator for Korean, despite pointing out in the comments that it is natural to omit pronouns in Korean sentences, repeatedly makes annotations for *Omitted Pronoun* where they are not necessary, drastically increasing the count of wrong annotations for this language and influencing the overall percentage as well.

In the case of Japanese, as seen in **Table 1**, the percentage of annotation mistakes is almost the same across the three categories. Through detailed analysis of the annotations, it becomes clear that for this language most mistakes are related to either the typology lacking appropriate categories or the annotation guidelines being unclear. This can be verified not only in terms of wrong annotations but also wrong span selection, specifically in annotation of verb related errors and omission, which is due to the fact that Japanese does not use whitespaces.

1	lp	source	target	typology_error	error	severity	wrong annotation	wrong span	wrong severity
31	en_ja	Would you be able to try accessing the academy course on a Guest profile?	ゲストプロフィールでアカデミーコースにアクセスしていただけますか。	mt_hallucination	リ	minor	yes	yes	yes

Figure 8. Example of wrong annotations in Japanese

The example on **Figure 8** shows a segment translated into Japanese where the annotation is wrong in all the required variables. The error identified by the annotator is a misspelling of the word for “academy”, since one of the characters in the translation is wrong.

However, as per the Unbabel Error Typology, the *Spelling* category is not selectable, as it branches out into two other issue types - *Source/Target Disagreement* and *Wrong Paronym*. Looking into the definition the guidelines provide for each of these issue types, it becomes easy to understand why the annotator avoided using any of them. According to the definition provided in the guidelines, a *Source/Target Disagreement* error occurs when “there are gender/number mismatches between source and target”, while the definition for *Wrong Paronym* states that “the target text has a paronym (a word written in a similar way to another word), and this results in a structure with a completely different meaning”. None of these error types is entirely compatible with the error at hand, as it is not an error of gender or number and the resulting word also does not have any meaning in Japanese. However, this typology also contains an *Orthography* issue type which is defined as “words spelled incorrectly” in the annotation guidelines. Despite this, the annotator considered that the error was an *MT Hallucination*, since the definition for this error is also relatively looser: “Content of the target text does not match with the content on the source text because of the machine translation”, revealing the annotation guidelines are being misinterpreted in some cases. As such, due to the fact that *Orthography* was the correct issue type to use for this error the annotation was considered to be wrong. The reason why the span selection for this error is also wrong is that the annotator tagged the error solely on the character that is wrong instead of selecting the whole word, as shown in **Table 2**. On the other hand, the severity attribution is wrong as well due to the fact that this is an error that can confuse the end user, as it results in a word that does not exist. Taking this into consideration, the error should have been tagged as major.

Source Text	Target annotated with wrong span and severity	Target annotated with correct span and severity	Target Text (corrected)
Would you be able to try accessing the academy course on a Guest profile?	ゲストプロフィールでアカデ リ ーコースにアクセスしてみただけですか。	ゲストプロフィールでアカデ リ ーコースにアクセスしてみただけですか。	ゲストプロフィールでアカデ ミ ーコースにアクセスしてみただけですか。

Table 2. Wrong span annotation in Japanese

Finally, it can be also concluded that although attribution of severities is where fewer mistakes happen with Japanese and Korean, it is still a relevant issue for all LPs under analysis. Furthermore, from an overall point of view, attribution of inadequate severities seems to be a bigger problem than wrong span selection, which is also problematic in the sense that, as mentioned before, severity influences MQM scores. This was taken into account when creating the guidelines for the annotation module discussed in **Section 4.3.2.1.**, which contains a dedicated section and a decision tree to guide annotators in the process of choosing the right severity for each error.

From this analysis it was possible to take note of the negative impact that an inadequate typology and guidelines can have on the quality and consistency of annotations. As verified through the data that was analyzed, this can lead to an overall problem of agreement between annotators and under-annotation, as the lack of appropriate issue types often leads the annotators to ignore these errors or to arbitrarily decide which issue types they should use for errors that do not have a directly corresponding issue type, meaning that two annotators may choose different issue types for the same error depending on their opinion or that even the same annotator can use the same issue type to annotate the same mistake in separate occasions.

4.3.1.1. Missing Categories

As mentioned before, one of the most important factors in creating the annotation module proposed in this thesis was to address the issue types that did not exist in the Unbabel Error Typology and that can have an impact in the annotation of East Asian languages, due to being important components of the same. As such, this section will explain the new specific issue types that were introduced and that are thought to allow the improvement of annotation quality and consistency for the languages under discussion. When possible the new proposed issue types will also be compared to those existent in the Asian language-focused error typology proposed by Ye and Toral (2020) in order to understand where the two typologies overlap and what their differences are.

- **Particles:** The first categories that were added to the annotation module due to the fact that there was a high degree of inconsistency in annotation of these function words were two issue types related to particles, or postpositions.

New Issue Types - Particles					
ACCURACY	→	Omission	→	Omitted Particle	
LINGUISTIC CONVENTIONS	→	Grammar	→	Function Words	→ Wrong Particle

Table 3. New issue types related to particles

Before explaining the need for particle related issue types in the East Asian Languages Annotation Module for the Unbabel Quality Framework, it is necessary to define what particles are and what is their role in the languages the annotation module is going to cover. In Japanese:

“Particles are attached to nouns and other words or phrases to show their grammatical function and role within the sentence or phrase (e.g. topic, subject, direct or indirect object, etc.). They do not occur as independent words. Particles always come after the word, phrase, or clause to which they relate.” (Bunt, 2003)

Thus, the definition of particles in Japanese is compatible with the definition for Korean particles, which is as follows:

“Particles are words that mark grammatical relationships, focus, emphasis, attitude, and a variety of emotional meanings. A Korean particle follows the word or phrase which it is marking (...)” (Martin, 1992)

A comparison between these two definitions leads to the conclusion that the particles used in Japanese and Korean are very similar in their functions and usage. In the case of both languages, most of the particles understood under this category are case particles and, as further clarified by Masuoka (1987) (as cited in Chida, 2015) they are “markers that show the relationship between a noun phrase and predicate”. As Chida (2015) further explains, Japanese case particles are postpositions which are attached to words with the function of indicating the relationship between words in a sentence. The same is true for Korean postpositions, as the definition provided by studies on particles is compatible with the one quoted above. In Korean, particles are divided into case and discourse or modal particles and their usage can be simultaneous, meaning that more than one particle can be stacked together (Lee et al., 2009).

In addition to case particles, there are other types of particles in Korean and Japanese which are important to mention. Both in Japanese and Korean the existence of sentence-final or sentence end particles is recognized and, as their name indicates, these are particles that are positioned at the end of a sentence and serve the functions of marking the clause or determining the speech style or mood (Pak, 2008). In the case of Japanese, such particles may also serve to distinguish the sex of the speaker, as there are particles that are used exclusively by men or women (Makino & Tsutsui, 1989). Furthermore, in relation to Japanese, Akamatsu (2011) defends that the honorifics that can often be found before names, verbs, adjectives and adverbs should also be considered as honorific particles.

However, in terms of the annotation module, although a clear distinction is not made between types of particles, it is assumed that the new issue types related to particles that were introduced will not extend to sentence-final and honorific particles in Japanese and Korean due to the fact that the proposed annotation module, as well as the Unbabel Error Typology which

was a part of the analysis on this thesis, contain the issue types of *Register*²¹ and *Tense/Mood/Aspect*. Due to the fact that honorific particles are directly related to the formality level, or register, of the text, their misuse or omission is thought to be more intuitively annotated as *Register* errors. The case of sentence-final particles for Japanese and Korean is similar in the sense that they usually affect the tense, mood or aspect of the sentence or even the register as well, as seen in this example from Makino and Tsutsui (1989):

(1) is a declarative sentence. If the sentence-final particles *ka* and *ne* are affixed to (1), as in (2) and (3), the sentence becomes a question ((2)) and a sentence of confirmation ((3)).

- (1) 山田さんは先生です。
Yamada-san wa sensei desu.
(Mr. Yamada is a teacher.)
- (2) 山田さんは先生ですか。
Yamada-san wa sensei desu ka.
(Is Mr. Yamada a teacher?)
- (3) 山田さんは先生ですね。
Yamada-san wa sensei desu ne.
(Mr. Yamada is a teacher, isn't he?)

Figure 9. Sentence-final particles in Japanese extracted from Makino and Tsutsui (1989:45)

While the Chinese language does not possess particles that function in the same way as particles in Japanese and Korean as they are defined above, it does not mean that particles are not used in Chinese, but rather that they exist in a different context. When talking about particles in Chinese, studies and grammars are referring to sentence-final particles. As their title indicates, these are mostly positioned at the end of sentences, but they can also be used in other contexts and have more than one function (Li & Thompson, 2009). As further defined by Tang (2015), sentence-final particles, or utterance particles, are “functional words that occur in the

²¹ While the annotation module proposed in this thesis and the current version of the Unbabel Error Typology (v3) only contain one register issue type (*Register*), the version of the Unbabel Error Typology which was tested (v2) separates register errors into two issue types (*Lexical Register* and *Grammatical Register*).

sentence-final position, expressing some grammatical meanings and pragmatic information, such as the speaker’s attitude”.

Although, as explained before, Japanese and Korean also possess sentence-final particles, due to the difference in structure between those languages and Mandarin Chinese these particles are viewed differently. While in Japanese and Korean it seems to be more natural to annotate instances involving these particles as *Tense/Mood/Aspect* or *Register* errors, due to the fact that Mandarin Chinese, in opposition, is a language where words are not inflected, there is a more distinct separation of sentence-final particles in this case and it may be more natural to annotate them as particle errors. However, the guidelines for the annotation module purposefully do not specify what constitutes a particle or not for each of these languages, so it is ultimately in the annotators’ hands to decide if these issue types are adequate based on the definitions presented on **Tables 4** and **5**, as per the provided guidelines:

<p>Omitted Particle A particle is missing in the target text.</p> <p>Ex: (EN) If for any reason your exchange request is rejected, we will contact you as soon as possible. (JA) 何らかの理由で交換リクエストが却下された場合 [O]OMITTED PARTICLE、できるだけ早くご連絡いたします。 (JA) 何らかの理由で交換リクエストが却下された場合 [は]OMITTED PARTICLE、できるだけ早くご連絡いたします。</p>

Table 4. Definition of *Omitted Particle* errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework

Wrong Particle

A particle is used incorrectly (another particle should have been used instead).

Ex:

(EN) [PRODUCT] account's registered email.

(KO) [PRODUCT] 계정 **[의]** WRONG PARTICLE 등록된 이메일.

(KO) [PRODUCT] 계정 **[에]** WRONG PARTICLE 등록된 이메일.

Table 5. Definition of *Wrong Particle* errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework

It should be pointed out that the error typology proposed by Ye and Toral (2020) also contains issue types related to particles. As will be explained in more detail in **Section 4.3.2.**, in the case of Ye and Toral’s typology the *Particle* issue type is a daughter issue type of the *Extraneous* category, which is meant to be used for annotation of errors related to the abusive use of function words in the context of translation. However, because the two issue types for annotation of particles were introduced in the East Asian Languages Annotation Module specifically because they had been noted to cause annotation mistakes in Japanese and Korean, it is not expected that Chinese annotators make as much use of these issue types. Ultimately, however, how the annotators for all languages use these two issue types will serve as grounds to study ways of further improving the annotation module in the future.

- **Classifiers:** Following particles, it was necessary to add parallel omission and incorrect usage issue types for classifiers.

New Issue Types - Classifiers					
ACCURACY	→	Omission	→	Omitted Classifier	
LINGUISTIC CONVENTIONS	→	Grammar	→	Function Words	→ Wrong Classifier

Table 6. New issue types related to classifiers

As defined by Koo (1997), measure words, or classifiers, are “words or morphemes used to indicate that a word belongs to a particular class or category”. The use of measure words is common to all four languages analyzed in this work and they should be used before nouns that are being counted (Koo, 1997). As pointed out by Fang and Conelly (2008) and Unbabel’s Language Guidelines for Simplified Chinese²² available at the time the annotations for this data analysis were made²³, while these words are also used in other languages like English (i.e. a *piece* of paper), in Chinese their usage is more complex, as the correct measure word that should be used differs according to the category of the noun they relate to and they are indispensable when describing a countable unit, as in the example on **Table 7**.

English	a/one pen
Chinese	一支笔 ²⁴

Table 7. Classifiers in Chinese

In fact, the error typology proposed by Ye and Toral (2020), which was introduced in **Section 3.4.2.**, also contained an issue type for classifiers:

Ye and Toral’s Typology				
FLUENCY	→	Grammar	→	Classifier

Table 8. Classifier issue type in the typology proposed by Ye and Toral (2020)

²² Available at: <https://help.unbabel.com/hc/en-us/articles/360008780374-Language-Guidelines-Chinese-Simplified->

²³ This data analysis was conducted between September and December of 2021 and the annotation data analyzed corresponded to the second semester of 2021.

²⁴ 支 is the measure word used for small elongated objects (Fang & Conelly, 2008).

While errors related to classifiers were not prolific in the data that was analyzed, due to the fact that there is no indication in both the Japanese and Korean language guidelines provided by Unbabel, and that in the case of Simplified and Traditional Chinese these words are mentioned under “Determiners” while still referring to them as particles and recognizing that it is not usual to use determiners in Chinese, it is not clear to annotators how they should be annotated.

As such, it was considered that including issue types concerning classifiers would be beneficial, not only to make the annotation process as easy and intuitive as possible for the annotators but, in consequence, to also improve the agreement between annotators by removing the necessity for subjective decisions. **Tables 9** and **10** illustrate the definition and examples provided for each issue type:

<p>Omitted Classifier A classifier is missing in the target text.</p> <p>Ex: (EN) Thank you for choosing [PRODUCT]. This is to inform you that you have opened a duplicate ticket. (ZH-CN) 感谢您选择 [PRODUCT], 谨此通知您, 您已发起了[Ø]OMITTED CLASSIFIER重复的查询。 (ZH-CN) 感谢您选择 [PRODUCT], 谨此通知您, 您已发起了一项]OMITTED CLASSIFIER重复的查询。</p>
--

Table 9. Definition of *Omitted Classifier* errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework

Wrong Classifier

A classifier is used incorrectly (another classifier should have been used instead).

Ex:

(EN) Click the New Account pop-up menu, then choose a type of user.

(ZH-CN) 点击新账户弹出菜单，然后选择~~一个~~**WRONG CLASSIFIER**类型的用户。

(ZH-CN) 点击新账户弹出菜单，然后选择~~一种~~**WRONG CLASSIFIER**类型的用户。

Table 10. Definition of *Wrong Classifier* errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework

- **Transliteration:** The fifth and final new issue type introduced in this annotation module is *Transliteration*.

New Issue Type - Transliteration			
ACCURACY	→	Mistranslation	→ Transliteration

Table 11. New issue type for transliteration

As defined by the Cambridge English Dictionary, transliteration is “the act or process of writing words using a different alphabet” (*TRANSLITERATION | Meaning in the Cambridge English Dictionary, 2022*). In the specific case of this annotation module transliteration should be understood as the act of writing words from the English source using characters in the target language, forming words that are phonetically similar to the original language. While there are several cases where the transliterated form of a word is the correct translation due to it being a loanword, transliteration can also occur in other instances, producing target words that either do not exist or are not appropriate to use depending on the context.

The need for the inclusion of this issue type in the annotation module was verified when the analysis of previous annotated data revealed that annotators were not clear on how to identify these errors, meaning that *Transliteration* errors were not consensually annotated and various issue types such as *Overly Literal* and *Lexical Selection* were used, as seen in the **examples 11** and **12**.

Annotation 1:

EN (source)

(11a) That Email will **assist** on how to create a new password for your account.

JA (target)

(11b) そちらのメールが、お客様のアカウントの新しいパスワードを作成する方法についてアシストいたします。

[Overly Literal, minor]

Annotation 2:

EN (source)

(12a) We will also follow up with Ikuo to **finalize** our approach to fixing this issue.

JA (target)

(12b) また、この問題の修正をファイナライズするため、Ikuoにフォローアップいたします。

[Lexical Selection, major]

The definition for *Transliteration* and examples provided in the guidelines are as follows:

Transliteration

A term in the target has been transliterated instead of being accurately translated.

Ex:

(EN) The knob on the front is for your headphone's direct monitoring/master volume.

(KO) 전면의 [노브]TRANSLITERATION는 헤드폰의 직접 모니터링/마스터 볼륨을 위한 것입니다.

(KO) 전면의 [손잡]TRANSLITERATION는 헤드폰의 직접 모니터링/마스터 볼륨을 위한 것입니다.

Table 12. Definition of *Transliteration* errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework

The annotation of the errors the five issue types described in this section are meant to cover has been inconsistent due to the fact that the typology previously in use at Unbabel did not include appropriate issue types and corresponding guidelines for these errors, leading the annotators to make decisions that were more subjective than desired. As such, the purpose of adding these issue types to the East Asian Languages Annotation Module for the Unbabel Quality Framework was to unify the annotation of the errors related to these categories, reducing the number of wrong annotations and increasing IAA scores and the consistency of annotations.

As can be seen in the table on **Annex A** (not shown here due to its length), the five new issue types introduced in this section are exclusive both in relation to the Unbabel Error Typology and the typology proposed by Ye and Toral (2020). However, it should be noted that, excluding *Transliteration*, Ye and Toral's typology contains issue types that can be considered almost equivalent, but differ in their specificity and organization. Where the East Asian Languages Annotation Module for the Unbabel Quality Framework proposes the two distinct issues type of *Omitted Particle* and *Wrong Particle*, Ye and Toral propose the *Missing* and *Incorrect* issue types to annotate these errors in relation to all function words. Similarly, the East Asian Languages Annotation Module proposes the issue types of *Omitted Classifier* and *Wrong Classifier*, while Ye and Toral's typology contains a single *Classifier* issue type, which is meant to be used in the case of incorrect usage of classifiers and it is assumed that, based on this definition, their omission should be annotated using the general *Omission* issue type in the typology.

4.3.1.2. Tests on Usability

Due to the fact that both the data analyzed in **Section 4.3.1.** and the annotations that were evaluated in order to provide feedback to annotators demonstrated that annotators of Japanese and Korean had considerable difficulty with the selection of the right span for each error, particularly concerning annotation around verbs and non-existent whitespaces, it was important to verify whether this was exclusively related to the lack of specific rules in the guidelines or to

the usability of the Annotation Tool. It was thought that the manual process of selecting errors in the Annotation Tool could present issues depending on the device or browser used to access it. As such, an investigation was conducted using several different pairings of devices and browsers in order to determine which, if any, resulted in difficulties regarding span selection.

For the purposes of this investigation, the batch that was created was to be annotated using the four language pairs under discussion on this thesis and the following five pairings of devices and browsers:

- 1) MacBook/Google Chrome
- 2) PC/Microsoft Edge
- 3) PC/Mozilla Firefox
- 4) Android phone/Google Chrome
- 5) iPad/Safari

For the pairings from 1 to 3 the annotations were done successfully. However, annotating on mobile devices proved to be extremely difficult or almost impossible so the batches were not annotated using pairings 4 and 5 and it was assumed that annotators do not use them either.

Contrary to what was initially thought and the motivation behind this investigation, at the time these annotations were performed, apart from a few very minor design issues when annotating using Microsoft Edge, there were no issues with the selection of errors for these languages in any of the devices tested. This led to the conclusion that the reason for having a considerable percentage of span selection errors is that there is not enough guidance in the guidelines concerning specific issues that arise when annotating these languages and that it was necessary to address them carefully when constructing the guidelines for the annotation module, which will be explained in **Section 4.3.2.**

4.3.2. The East Asian Languages Annotation Module for the Unbabel Quality Framework

The East Asian Languages Annotation Module for the Unbabel Quality Framework, represented on **Figure 10**, was built based on the version of the Unbabel Error Typology which

has been in use as of June of 2022 (v3). It must be highlighted that this is not the same version discussed in **Section 3.4.1.** but rather a new adaptation that was built in sequence of the integration with Lingo24, in order to be more inclusive and appropriate for the newly supported content and types of translation. However, due to time constraint reasons it has not been possible to perform a comparison between that typology and the annotation module proposed in this work.

The East Asian Languages Annotation Module for the Unbabel Quality Framework

The East Asian Languages Annotation Module for the Unbabel Quality Framework proposed in this thesis was built to attempt to resolve the problems in annotation that were verified in the data analyzed, which was annotated using the version of the Unbabel Error Typology that was in use until June of 2022 (v2). At the same time, as the ultimate objective is to be able to put this annotation module in production for it to be used effectively at the company for annotation of Japanese, Korean, Simplified and Traditional Chinese content, it would not make sense to diverge much from the newest version of the Unbabel Error Typology (v3), which is being currently used. As such, this annotation module was built with strong basis on newest version of the Unbabel Error Typology (launched after June 2022) while still maintaining many of the features from the previous version in order to carry out a comparison that would allow to see if the East Asian Languages Annotation Module for the Unbabel Quality Framework is efficient in solving the problems that were previously noted.

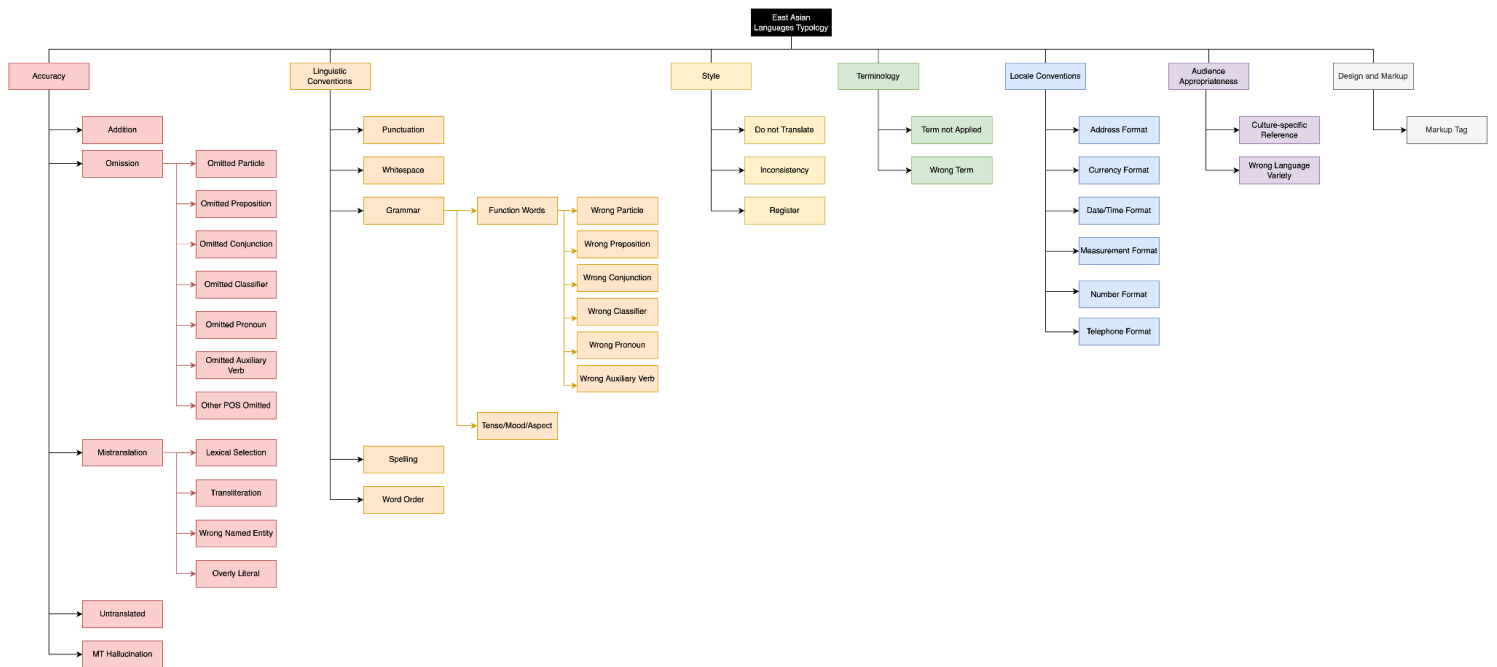


Figure 10. East Asian Languages Annotation Module for the Unbabel Quality Framework

The East Asian Languages Annotation Module for the Unbabel Quality Framework, represented on **Figure 10**, contains a total of 7 coarse categories, 24 daughter issue types, 13 granddaughter issue types and 6 great-granddaughter issue types, amounting to a total of 39 selectable issue types.

As mentioned above, the distribution of the issue types per category and overall structure of the annotation module attempt to respect the newest version of the Unbabel Error Typology (v3). However, there are a few characteristics that are notably not the same between the two typologies and where the East Asian Languages Annotation Module for the Unbabel Quality Framework is closer to the previous version of the Unbabel Error Typology (v2). As such, it is important to explore what those differences are and why they exist.

- Omission

In opposition to the newest version of the Unbabel Error Typology (v3), where *Omission* is a single issue type with no further distinctions, the East Asian Languages Annotation Module

for the Unbabel Quality Framework separates *Omission* into seven issue types depending on the type of error, much in the same way as in the previous version of the Unbabel Error Typology (v2). The issue types under *Omission* for the East Asian Languages Annotation Module for the Unbabel Quality Framework are:

- **Omitted Particle:** a particle is missing in the target text;
- **Omitted Preposition:** a preposition is missing in the target text;
- **Omitted Conjunction:** a conjunction is missing in the target text;
- **Omitted Classifier:** a classifier is missing in the target text;
- **Omitted Pronoun:** a pronoun is missing in the target text;
- **Omitted Auxiliary Verb:** an omitted auxiliary verb is missing in the target text;
- **Other POS Omitted:** one or more words belonging to any morphological category are missing.²⁵

The choice to maintain the distinction within *Omission* issue types came from the fact that a substantial part of the errors found in previous annotation data were related to the different types of *Omission*, as seen in **Section 4.3.1.1.** where the reason for introducing *Omitted Particle* and *Omitted Classifier* issue types was justified. As the objective of this work is both to create a new annotation module that suits East Asian languages and to study its impact, it was important to maintain and improve the key problem areas in order to have a better comparison.

On the other hand, as can be verified in the table on **Annex A**, the issue type *Omitted Determiner* was not transferred from the Unbabel Error Typology to the East Asian Languages Annotation Module. This is due to the fact that in previous annotation data this issue type was seldom used across all four language pairs and, when it was applied, it was always wrongly done so. Additionally, as it will be explained later in **Section 5.1.3.**, this was also an issue type that was identified in previous annotator feedback for these languages as unnecessary.

²⁵ The definitions for each issue type in this section are the same definitions used in the Annotation Guidelines for the East Asian Languages Annotation Module for the Unbabel Quality Framework which were provided to the annotators. The full version of the guidelines can be consulted in the Annexes of this thesis.

- Mistranslation

Another problem area of the annotation data from the previous typology was *Mistranslation*, which was maintained in the East Asian Languages Annotation Module for the Unbabel Quality Framework as a parent issue type with corresponding daughter issue types in similarity to that version (v2) of the typology. However, there was an attempt to balance the way *Mistranslation* was used in that version (v2)²⁶ and the way it exists in the current typology, which is as a single issue type, with the addition of the *Transliteration* issue type discussed in **Section 4.3.1.1** and removal of issue types from the previous version (v2) which were considered too ambiguous and identified as unnecessary by the annotators in previous feedback, such as *Ambiguous Translation* and *False Friend*. In light of this, the final issue types under *Mistranslation* in the East Asian Languages Annotation Module for the Unbabel Quality Framework are as follows:

- **Lexical Selection:** the term selected is not correct in context or does not accurately convey the meaning of the original text;
- **Transliteration:** a term in the target has been transliterated instead of being accurately translated;
- **Wrong Named Entity:** the target contains an error related to named entities (names, places, etc.);
- **Overly Literal:** the translation is too strict to the source text, which results in problems of interpretation (like literal translations of idiomatic expressions).

- Grammar

Similarly to *Omission* and *Mistranslation*, *Grammar* is an issue type that from the previous Unbabel Error Typology to the newest version currently in use (v3) lost its daughter issue types. However, in the East Asian Languages Annotation Module for the Unbabel Quality Framework this is a category that continues to be divided. This is due to the fact that *Grammar* is a very vague error type in itself and would not allow to properly evaluate the effect of

²⁶ In the previous version of the Unbabel Error Typology (v2), *Mistranslation* contained 10 daughter issue types and 2 granddaughter issue types.

introducing specific new issue types, such as *Wrong Particle* and *Wrong Classifier*, and specific guidelines on how to annotate *Tense/Mood/Aspect* errors, since these were also issue types where many annotation errors occurred. Similarly to what was verified with *Omission*, the issue type related to determiners which existed under the Unbabel Error Typology was also removed from the *Function Words* category in the East Asian Languages Annotation Module due to it not being used frequently in previous annotation data and being incorrectly applied in the few instances the annotators selected this issue type.

As such, the *Grammar* issue type in the East Asian Languages Annotation Module for the Unbabel Quality Framework is further distributed in the following manner:

- **Function Words:**
 - **Wrong Particle:** a particle is used incorrectly (another particle should have been used instead);
 - **Wrong Preposition:** a preposition is used incorrectly (another preposition should have been used instead);
 - **Wrong Conjunction:** a conjunction is used incorrectly (another conjunction should have been used instead);
 - **Wrong Classifier:** a classifier is used incorrectly (another classifier should have been used instead);
 - **Wrong Pronoun:** a pronoun is used incorrectly (another pronoun should have been used instead);
 - **Wrong Auxiliary Verb:** an auxiliary verb is used incorrectly (another auxiliary verb should have been used instead);
- **Tense/Mood/Aspect:** a verbal form displays the wrong tense, mood, or aspect.

As mentioned in **Section 3.4.2.**, the error typology proposed by Ye and Toral (2020) is directed specifically at a language pair that the annotation module this work includes, namely the translation direction of English to Chinese. As such, it is of interest to compare how both taxonomies approach *Grammar*, as the results end up being quite different.

In Ye and Toral's typology *Grammar* contains a total of seven selectable issue types, where five correspond to function words related errors as follows:

- **Extraneous:**
 - **Preposition:** issues related to the excessive use of prepositions;
 - **Adverb:** issues related to the excessive use of adverbs;
 - **Particle:** issues related to the excessive use of particles;
- **Incorrect:** a function word is used incorrectly in the target text;
- **Missing:** a function word is missing in the target text.²⁷

Ye and Toral define *Extraneous* errors as those that correspond to westernized Chinese expressions. As stated in Tse (2001) (as cited in Ye & Toral, 2020), “westernized Chinese refers to a cross-lingual phenomenon of imposing English grammar on Chinese, which is manifested in many problematic forms, abuse of function words especially”. This category is exclusive to the typology proposed by Ye and Toral, as it does not exist in the taxonomy proposed in this thesis. Its usage, however, will be analyzed in the **Results** section of this work in order to verify if, in case it was used by the annotators, it was correctly applied. As mentioned in **Section 3.4.2.**, guidelines for Ye and Toral's typology were created with basis on their work and MQM error definitions and then provided to the annotators. As per the guidelines, definitions and examples for *Extraneous* errors are as follows:

²⁷ The definitions provided for these issue types are the same as in the guidelines which were created for Ye and Toral's typology in order to assist the annotators in the testing phase of this work.

<p>Extraneous</p> <p>Westernized expressions. This refers to the phenomenon where English grammar is imposed on target sentences, which can be manifested especially in the abuse of function words.</p>	<p>Preposition</p> <p>Issues related to the excessive use of prepositions.</p> <p>Ex:</p> <p>(EN) Make sure that the audio is not set to "mute" on the audio source.</p> <p>(ZH-TW) 確保聲音未在音源上 [被]EXTRANEOUS PREPOSITION 設置為「靜音」。</p> <p>(ZH-TW) 確保聲音未在音源上 [Ø]EXTRANEOUS PREPOSITION 設置為「靜音」。</p>
	<p>Adverb</p> <p>Issues related to the excessive use of adverbs.</p> <p>Ex:</p> <p>(EN) If you are unsure which version you have, you can always uninstall the app and reinstall to make sure you have the most recent version.</p> <p>(JA) お持ちのバージョンがわからない場合は、[いつでも]EXTRANEOUS ADVERB アプリをアンインストールして再インストールして、最新バージョンであることを確認することができます。</p> <p>(JA) お持ちのバージョンがわからない場合は、[Ø]EXTRANEOUS ADVERB アプリをアンインストールして再インストールして、最新バージョンであることを確認することができます。</p>
	<p>Particle</p> <p>Issues related to the excessive use of particles.</p> <p>Ex:</p> <p>(EN) They are trying to release an update weekly until the problem is fixed.</p> <p>(JA) 問題が修正されるまで、毎週 [の]EXTRANEOUS PARTICLE アップデートをリリースできるよう取り組んでおります。</p> <p>(JA) 問題が修正されるまで、毎週 [Ø]EXTRANEOUS PARTICLE アップデートをリリースできるよう取り組んでおります。</p>

Table 13. Definition of *Extraneous* errors in the guidelines created for the typology proposed by Ye and Toral (2020)

While *Extraneous* errors are not a part of the East Asian Languages Annotation Module for the Unbabel Quality Framework, *Incorrect* and *Missing* have parallels in *Function Words* and *Omission*, as represented in **Table 14**.

Error typology proposed by Ye and Toral (2020)	East Asian Languages Annotation Module
Incorrect	<u>Function Words</u> : Wrong Particle, Wrong Preposition, Wrong Conjunction, Wrong Classifier, Wrong Pronoun, Wrong Auxiliary Verb
Missing	<u>Omission</u> : Omitted Particle, Omitted Preposition, Omitted Conjunction, Omitted Classifier, Omitted Pronoun, Omitted Auxiliary Verb, Other POS Omitted

Table 14. Comparison of issue types between the typology proposed by Ye and Toral (2020) and the East Asian Languages Annotation Module for the Unbabel Quality Framework

As mentioned before, due to the fact that the East Asian Languages Annotation Module for the Unbabel Quality Framework was meant to be as similar as possible to the ones in use at Unbabel, the distribution of these errors was maintained under *Function Words (Grammar)* and *Omission*, instead of joined together as in the typology proposed by Ye and Toral. Whether this makes a difference in how the annotators perceive each error type will be analyzed in **Section 5**.

Finally, it is important to mention how *Tense/Mood/Aspect* is an issue type that, out of these two typologies, is only a part of the East Asian Languages Annotation Module for the Unbabel Quality Framework. As stated by Ye and Toral (2020): “the relations between sentence parts, tenses and aspects are often shown through word order, particles or context in Chinese, due to its lack of inflection”. For this reason, their proposed typology does not include a *Tense/Mood/Aspect* issue type, as according to the definitions provided such errors could be annotated using the issue types under *Function Words*. However, due to the fact that the East Asian Languages Annotation Module for the Unbabel Quality Framework was created to also allow annotation of Japanese and Korean, it was essential to maintain this error type. Furthermore, it was believed that verbs in Chinese could also be annotated using this issue type, provided the guidelines were appropriate and clear in this regard. In light of this, the definition

for *Tense/Mood/Aspect* in the East Asian Languages Annotation Module for the Unbabel Quality Framework is as follows:

<p>Tense/Mood/Aspect</p> <p>A verbal form displays the wrong tense, mood, or aspect. Please include the entire span of the verb and do not consider the radical of the verb and its conjugation separately. In the case of Chinese, please use this tag to annotate all the components that affect the verb using the multi-selection function. For a more detailed explanation refer to the Tricky Cases section.</p> <p>Ex:</p> <p>(EN) We will not process your latest refund request.</p> <p>(KO) 고객님의 최신 환불 요청은 처리하지 [않습니다]<small>TENSE/MOOD/ASPECT</small></p> <p>(KO) 고객님의 최신 환불 요청은 처리하지 [않을 것입니다]<small>TENSE/MOOD/ASPECT</small></p>

Table 15. Definition of *Tense/Mood/Aspect* errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework

- Removed issue types

In the same way that a few specific issue types were especially added to the East Asian Languages Annotation Module for the Unbabel Quality Framework, some were also removed in relation to both versions of Unbabel’s typology with basis on one of two factors:

- The issue type is not applicable or relevant to the concerning language pairs;
- The issue type in the newest version of the Unbabel Error Typology (v3) is too different from any of the issue types in the previous version (v2) and would not be of value for a comparison.

An example of an issue type that is still a part of newest Unbabel Error Typology and was removed from the East Asian Languages Annotation Module for the Unbabel Quality Framework due to it not being applicable is *Capitalization*. As none of the four languages the annotation module is meant to cover uses the Roman alphabet as its writing system, capitalization is a phenomenon that can only occur in the case of untranslated named entities, such as usernames and product titles, cases in which the *Wrong Named Entity* issue type should

be used. In addition to avoiding unnecessary noise in the annotation taxonomy, this removal is thought to have the benefit of reducing the number of annotation mistakes, as in previous data annotators tended to use it incorrectly on *Named Entity* errors. In addition to *Capitalization*, other issue types under *Typography* in the Unbabel Error Typology were also removed in the East Asian Languages Annotation Module in accordance with previous annotator feedback, which can be consulted in **Tables 30** and **31** in **Section 5.1.3.** Such is the case of *Diacritics* and *Hyphenation*, which are not applicable to the four languages under discussion. Besides the issue types under *Typography*, the issue type of *Agreement* that had been previously under *Grammar* was also removed in relation to the Unbabel Error Typology due to it both not being applicable and having been pointed out by the annotators as unnecessary.

As mentioned before in this section, the fact that the newest Unbabel Error Typology aims to cover content brought in by Lingo24 that goes beyond customer service related content means that completely new issue types were also introduced, such as *Lacks Creativity* and *Unnatural Flow*. These issue types that differ greatly from the previous version of the typology (v2) were not included in the East Asian Languages Annotation Module for the Unbabel Quality Framework for the time being because they would not be applicable to the chat data that was to be annotated for the purposes of this thesis, and their inclusion would result in unnecessary noise and would not be of value at this stage. However, in future versions of the East Asian Languages Annotation Module for the Unbabel Quality Framework the introduction of these issue types will be reevaluated.

4.3.2.1. Guidelines

In order to help the annotators to fully understand the annotation module and to provide appropriate definitions for each issue type, a set of guidelines with examples and Decision Trees was also created. Although the full version of the guidelines will be available in the **Annexes**, it is pertinent to analyze in this section the specifications addressed under some issue types and the *Tricky Cases Section* which were all included in order to help eliminate ambiguity and make the annotation process as clear as possible.

Apart from the five new issue types presented in **Section 4.3.1.1.** and for which definitions and corresponding examples were already discussed, there are issue types that, based on what was noted during the data analysis phase of this work, needed further clarification.

As noted in the analyzed datasets, the MT systems often struggle with applying punctuation in their correct width in Chinese and Japanese text, producing many incidents of half-width punctuation being used incorrectly. However, even though this type of error was extremely common, the annotators usually do not make note of it and, as such, further specification regarding this issue was included in the guidelines. As such, the definition provided for Punctuation is as follows:

Punctuation

Punctuation is used incorrectly or is missing, or one of a pair of quotes, brackets or punctuation is missing from the target text. **Please note** that punctuation that is supposed to be full-width, as is the case for commas in Chinese and Japanese, should also be annotated using this tag if they don't appear as such in the target.

Ex:

(EN) Thank you for providing a screenshot.

(KO) 스크린샷[.]PUNCTUATION를 제공해 주셔서 감사합니다.

(KO) 스크린샷[Ø]PUNCTUATION를 제공해 주셔서 감사합니다.

Table 16. Definition of *Punctuation* errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework

The issue type of *Tense/Mood/Aspect* is another one which deserved further clarification both in terms of its definition and in the *Tricky Cases Section*, as seen in **Table 17.**

Due to the fact that in Chinese verbs are not inflected, they depend on function words to express tense. From the data analyzed previous to the creation of the annotation module, it was possible to conclude that due to this annotators were often confused on how to annotate verbs in Chinese, so there did not seem to be a consensus on how to do it, and they were forced to use more than one annotation to identify the same error, as seen in **example 13.**

EN (source)

(13a) In case of no reply, an automated reminder would be sent to you and if we still don't get any reply, this ticket will eventually be set as solved.

ZH-CN (target)

(13b) 如果您没有回复, 自动发送的提醒将发送给您, 如果我们还没有收到任何回复, 查询将最终被设置为解决。

[Other POS Omitted, major]

(13c) 如果您没有回复, 自动发送的提醒将发送给您, 如果我们还没有收到任何回复, 查询将最终被设置为解决。

[Tense/Mood/Aspect, major]

In **example 13** the same error was annotated using two issue types due to the fact that, according to the annotator's comment, the adverb 已 is missing before the selected word. In the same way, the opposite situation is also possible if an extra word is added that affects the tense, mood or aspect of the verb incorrectly. As such, it is proposed that while using this annotation module the annotators should make use of the multi-selection feature in the Annotation Tool to select all components that affect the tense, mood or aspect of the verb the error falls upon in order to both unify annotations and avoid multiple annotations for one error.

In addition to further explanation under the definition for each issue type, the guidelines also include a *Tricky Cases Section* which was created to address the differences between the categories which were thought could be more ambiguous and the annotation problems which were verified in the analysis phase of this work. For each of the issues addressed in this section a brief explanation or disambiguation statement was presented, followed by illustrative examples as seen in **Table 17**.

Annotating verbs

Annotating verbs can be tricky since they may contain various components. When annotating Japanese and Korean, the verb radical and the verb form shouldn't be considered separately. As such, when errors such as *Tense/Mood/Aspect* or even *Lexical Selection* fall upon a verb, the annotation should look like this:

Source Segment	Target Segment	Correct Target Segment	Annotation
Do not <u>press Bluetooth button</u> yet.	Bluetoothボタンをまだ <u>押し</u> ないでください。	Bluetoothボタンをまだ <u>押さ</u> ないでください。	Bluetoothボタンをまだ <u>押し</u> ないでください。 TENSE/MOOD/ASPECT

✗ Bluetoothボタンをまだ押しないでください。

✓ Bluetoothボタンをまだ押さないでください。

Source Segment	Target Segment	Correct Target Segment	Annotation
We will not process your latest refund request.	고객님의 최신 환불 요청은 처리하지 <u>않</u> 습니다.	고객님의 최신 환불 요청은 처리하지 <u>않을</u> 것입니다.	고객님의 최신 환불 요청은 처리하지 <u>않</u> 습니다. TENSE/MOOD/ASPECT

✗ 고객님의 최신 환불 요청은 처리하지 않습니다.

✓ 고객님의 최신 환불 요청은 처리하지 않을입니다.

In the case of Chinese, due to the lack of direct conjugation of the verbs, there can be two different situations:

- If the error falls upon the tense, mood or aspect of the verb, all the elements that make up the error should be selected. In the case of these elements being separated in the sentence, they should still be selected using the multi-selection function that is used for *Word Order* and *Inconsistency* errors:

Source Segment	Target Segment	Correct Target Segment	Annotation
We are not influencing your stats and cannot give you only good or bad teams.	我们不[影响]你的数据[了], 也无法只为你提供优秀或不良的团队。	我们不影响你的数据[Ø], 也无法只为你提供优秀或不良的团队。	我们不影响你的数据[了], 也无法只为你提供优秀或不良的团队。 TENSE/MOOD/ASPECT
Note: In cases like this when an extra element in the sentence affects its tense, mood or aspect, please use this error type instead of others like Addition. Likewise, if there is a component missing that also affects the tense, mood or aspect of the sentence, instead of using omission tags please use Tense/Mood/Aspect.			

- If the error is a matter of *Lexical Selection*, only the unit containing the error should be selected:

Source Segment	Target Segment	Correct Target Segment	Annotation
I've checked and you haven't added any goals on the campaigns, that's why the reporting table is empty on those fields.	我已经[检查]过, 你没有在推广计划上添加目标, 因此报告表格中这些字段是空的。	我已经[查看]过, 你没有在推广计划上添加目标, 因此报告表格中这些字段是空的。	我已经检查过, 你没有在推广计划上添加目标, 因此报告表格中这些字段是空的。 LEXICAL SELECTION

Table 17. Tricky Cases section on annotating verbs with the East Asian Languages Annotation Module for the Unbabel Quality Framework

Finally, in addition to the Tricky Cases section and the explanations under the issue types, the guidelines also include the two decision trees represented in **Figure 11** and **Figure 12** to further help the annotators choose the right category and severity for each annotation respectively.

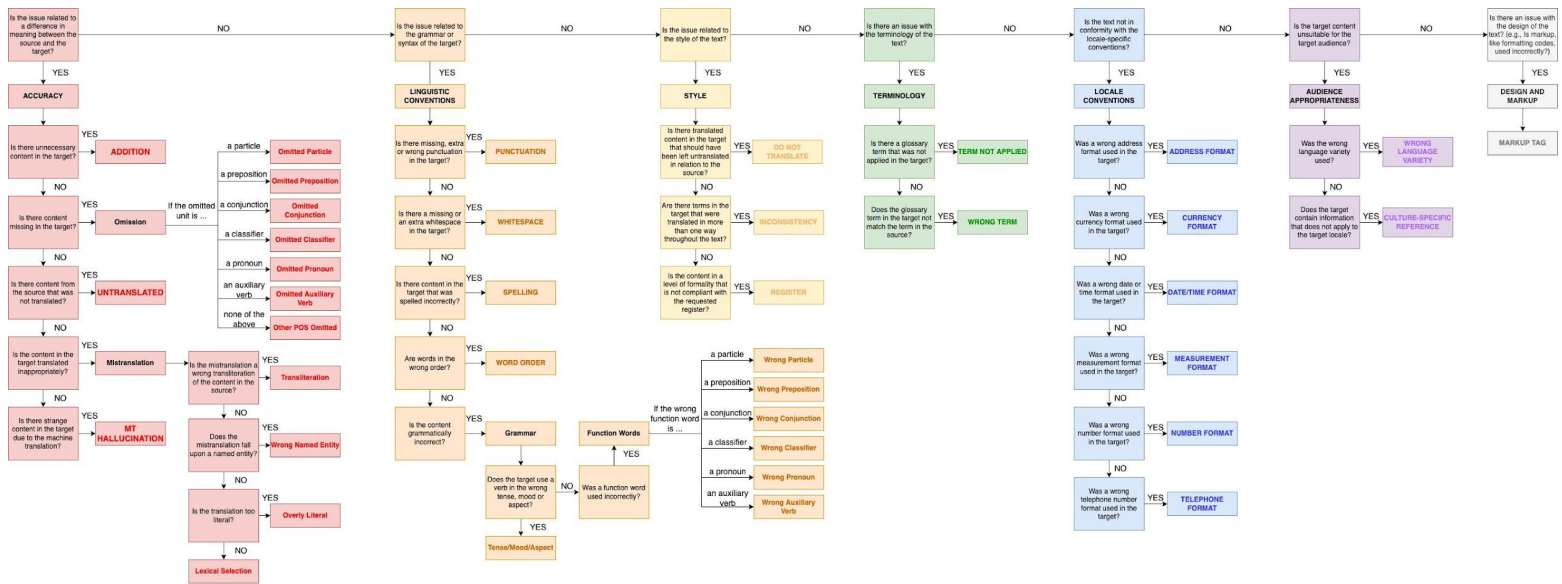


Figure 11. Decision Tree for issue type selection in the East Asian Languages Annotation
Module for the Unbabel Quality Framework

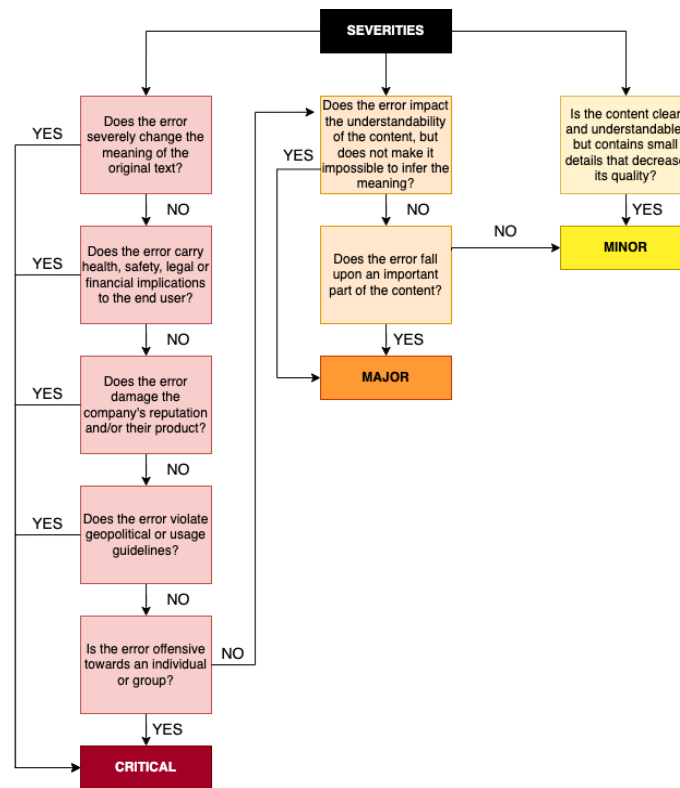


Figure 12. Decision Tree for severity selection in the East Asian Languages Annotation Module for the Unbabel Quality Framework

4.3.2.2. Annotator Training

Aside from the specific guidelines created for this annotation module, as well as the guidelines corresponding to Ye and Toral’s typology that was discussed previously, further annotator training was conducted to allow the annotation process to be as smooth as possible in order to obtain the best possible results. As such, during the annotation process all the annotators were allowed and encouraged to ask questions related to annotation difficulties with the two new typologies they had never worked with - the typology proposed by Ye and Toral (2020) and the annotation module proposed in this thesis.

The typology that generated the most questions was Ye and Toral’s. This was not unexpected as this taxonomy was considerably different from the one the annotators had been

using so far, which caused a high degree of unfamiliarity. Another reason for this was the fact that the Unbabel Error Typology was extremely Customer Service-oriented and, as such, contained many issue types adequate for annotation of that content. On the other hand, Ye and Toral's typology is both much more generalized in terms of content type and is oriented to one specific translation direction. For this reason it was only natural that the annotators required more training to be able to use this typology. Below is an adapted list of the most common questions the annotators asked regarding this typology:

- 1) How do I annotate a lexically mistranslated non-function word?
- 2) How do I annotate register errors?
- 3) How do I annotate whitespace errors?
- 4) How do I annotate terminology errors?
- 5) How do I annotate wrong language variety errors?

The instructions related to these questions were provided while keeping in mind both the limitations of the typology itself and the Annotation Tool and that the annotators should not be influenced to choose one determined issue type, as this would manipulate results and be unhelpful in studying the effectiveness of the typologies.

Regarding question number (1), it was suggested that the annotators should still annotate these errors while choosing which issue type from the typology as a whole they felt was more adequate. This question came from the fact that the only selectable issue types under Mistranslation in Ye and Toral's typology are *Overly-literal* and *Entity*. Due to the limitations of the Annotation Tool, and also because there was no specification regarding whether or not *Mistranslation* as a parent issue should be selectable or not, *Mistranslation* was ultimately not selectable. This created confusion due to the fact that the annotators felt that *Overly-literal* and *Entity* were not sufficient and the consequences of this will be analyzed in **Section 5**.

In relation to questions (2) and (3), the annotators were instructed to ignore these errors while annotating with this typology, since there were no issue types that were similar. However, it was recognized that this would influence IAA positively since it would remove two issue types the annotators were already overlooking inconsistently in previously analyzed data, so their removal altogether meant no one would be using these error types regardless and agreement

would go up. As such, this was noted as something to keep in mind when analyzing the annotation results.

Finally, it was recognized that questions (4) and (5) came up due to the specificity of the annotated content in contrast with the typology itself. In other words, this typology was created with the intent to serve only the translation direction of English to Chinese and is not a specialized typology, hence its application to specialized content and more than one language variety, if we consider it is meant to be applied to only one variety of Chinese, goes beyond the original scope of the typology. As such, the annotators were once again advised to annotate these errors with the issue type they felt was closest.

Regarding the East Asian Languages Annotation Module for the Unbabel Quality Framework, the only question asked by the annotators was regarding further clarification on the *Tricky Cases Section* about annotation of particles.

In addition, one of the annotators surprisingly asked a question about the previous version of the Unbabel Error Typology (v2), which was used for the first round of annotations which will be analyzed in the following section. Interestingly enough, this question was on how to suggest new issue types missing from the typology, such as issues related to Classifiers, as pointed out by the annotator. In this case the annotator was advised to annotate these errors as they had previously while using the Unbabel Error Typology.

This training was essential in clearing up annotators' doubts about the new typologies and it was something all annotators participated in with pertinent questions and remarks about the typologies, proving useful for analyzing the annotation results and gathering data on how to further improve the East Asian Languages Annotation Module for the Unbabel Quality Framework in the future.

This chapter discussed the objectives of this internship and the work that was developed during its duration. In addition to a brief description of the tasks performed at the company during the integration with Lingo24, the process of building the East Asian Languages Annotation Module for the Unbabel Quality Framework was also presented, which included the identification of existing errors and the reasons behind those issues. Finally, based on the data

obtained during this analysis, the East Asian Languages Annotation Module for the Unbabel Quality Framework was created together with its respective guidelines, which are presented in the last section of this chapter.

5. Results and Discussion

After building the East Asian Languages Annotation Module for the Unbabel Quality Framework and its respective guidelines it was necessary to test whether this module can improve the quality of annotations for Japanese, Korean, Simplified and Traditional Chinese within Unbabel. In addition, it was important to determine the strengths and weaknesses of the annotation module by comparing it with other typologies and by evaluating how each annotator performed by looking directly at how annotations were made in the Annotation Tool.

For the purpose of obtaining a valuable comparison, a total of around 13,000 words were annotated across all of the four language pairs. Every dataset was annotated by two annotators, henceforth referred to as *Annotator A* and *Annotator B*, using all three of the following typologies: the Unbabel Error Typology that had been in use until June of 2022, the error taxonomy proposed by Ye and Toral (2020) and finally the East Asian Languages Annotation Module for the Unbabel Quality Framework, proposed in the context of this project.

In terms of annotation setup, although there was an attempt to have datasets as similar as possible in terms of number of jobs and translated words, there ended up being some disparities. The Japanese dataset was the shortest, with 14 jobs and a total of around 1100 words due to some limitations in terms of data availability. Both the Traditional and Simplified Chinese datasets were composed of 20 jobs each and a total of around 2400 and 4900 words respectively. Finally, the Korean dataset was the most extensive in terms of number of jobs (21) but not the one with the most words, having a total of 3700. The annotation process had the approximate duration of three months, from the end of April to the end of July 2022, with the jobs for the different typologies being distributed with one month intervals, which meant that the annotators had approximately a month to annotate with each typology.

In this section the results of this comparison will be discussed in relation to each of the error typologies and different aspects. Firstly, the IAA scores for each dataset will be presented and analyzed in contrast with some of the annotations that were made and the fluctuation of MQM scores. Secondly, the annotation choices of each annotator will be observed, particularly regarding issue types which were previously noted as being problematic. Finally, the feedback left by the annotators regarding each typology will also be discussed.

5.1. Unbabel Error Typology

The first typology used to annotate the provided dataset was the version of the Unbabel Error Typology which was explained in **Section 3.4.2.** and was in use until June of 2022 (v2).

5.1.1. Inter-annotator Agreement

For each of the four language pairs under discussion the IAA scores were compared in order to assess how efficient each typology was in solving ambiguities and, in consequence, allowing consistent annotations. A proper measurement of inter-annotator agreement implies that the annotators are following the same set of rules, communicated through guidelines, and that the annotators are independent from each other (Artstein, 2017). This ensures that the annotations are performed on the same basis and that annotators are not influenced by each other.

While it is true that using only one coefficient to measure IAA does not allow all aspects of annotation to be considered at once (Artstein, 2017), at the time of this experiment it was only possible to get IAA scores using the Cohen Kappa coefficient due to changes in the annotation platform. The Cohen Kappa coefficient is used to calculate the level of agreement between two raters, resulting in scores ranging from -1 to 1 in which a score of 1 represents perfect agreement. As seen in Amidei et al. (2019), based on previous studies, depending on the number of categories, the average IAA score obtained with this coefficient is 0.40, which corresponds to the fair to moderate rating proposed by Landis and Koch (1977) (as cited in Amidei et al., 2019). As such, for the purposes of IAA analysis on this thesis the threshold that defines an acceptable score was set at 0.4.

Table 18 represents the average IAA corresponding to the dataset annotated using the Unbabel Error Typology for each of the four language pairs.

Average batch IAA per Language Pair (LP)		
LP	Average batch IAA	% of jobs above 0.4 threshold
English-Japanese	0.628	54,5%
English-Korean	0.366	50%
English-Traditional Chinese	0.464	60%
English-Simplified Chinese	0.192	17,6%

Table 18. Average batch IAA per LP with the Unbabel Error Typology

From the results obtained it is possible to observe that while the Japanese and Traditional Chinese results fall above the threshold established as acceptable (0.4), both the Korean and Simplified Chinese results are below this threshold, with the Simplified Chinese dataset obtaining a particularly low score. In fact, as seen in **Table 18**, while at least half of the jobs for the other languages scored above the set threshold, in the case of Simplified Chinese only a small percentage of jobs obtained a score of 0.4 or more.

This disparity can be justified through a few reasons. Firstly, it was expected that the Japanese dataset would generate the most positive results of all language pairs due to the fact that the translations for Japanese have historically suffered less from chronic errors present in other languages such as Korean and Simplified Chinese.

In the case of Korean, the most common type of error across all jobs is the misuse of whitespaces. As Korean is a language that poses much difficulty in relation to the usage of whitespaces, this type of error occurs very often, whether it corresponds to the addition of an extra whitespace or to the lack of a whitespace between words. In the case of this dataset every job is very polluted with both types of mistakes. However, while *Annotator A* annotates every error, amounting to a large number of annotations in some jobs, *Annotator B* mostly ignores them, annotating only around 30% of the *Whitespace* errors *Annotator A* did, and generally only annotates instances of extra whitespaces, as seen in **Table 19**.

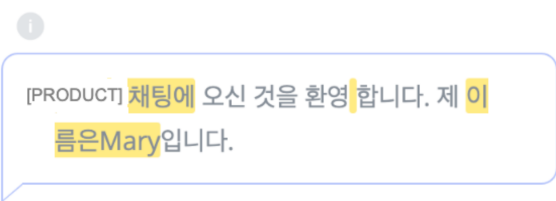

Annotation of whitespaces in Korean	
Annotator A	Annotator B
 <p>3 whitespace annotations</p>	

Table 19. Example of annotation of whitespaces in Korean with the Unbabel Error Typology

In the case of Simplified Chinese, as will be discussed in the following section, punctuation not being transformed into full-width in the target text was the origin of multiple annotations made by *Annotator A* and not corresponded by *Annotator B*, which affected IAA negatively.

However, this alone does not explain the low IAA scores obtained with this language pair. Apart from some disagreement in relation to issue types under *Mistranslation*, which will be explained in **examples 26 to 29**, the annotators for this language pair disagree on span, which results in the same errors being annotated differently, as seen in **examples 14 and 15**.

Span selection in Simplified Chinese	
Annotator A	Annotator B
<p>EN (source) (14a) it is already a combination of the 2.4 and 5ghz network since the device itself will be the one who will decide which network is the best for that specific device.</p> <p>ZH-CN (target) (14b) 它已经是 2.4 和 5GHz 网络的组合,因为该设备本身将是由谁决定哪个网络最适合该特定设备。 是由谁 [Addition, major]</p>	<p>EN (source) (15a) it is already a combination of the 2.4 and 5ghz network since the device itself will be the one who will decide which network is the best for that specific device.</p> <p>ZH-CN (target) (15b) 它已经是 2.4 和 5GHz 网络的组合,因为该设备本身将是由谁决定哪个网络最适合该特定设备。 是 [Addition, major] 由 [Addition, major] 谁 [Addition, major]</p>

Table 20. Example of span selection in Simplified Chinese with the Unbabel Error Typology

However, even though the Simplified Chinese dataset distinguishes itself with the worst average IAA scores, it is the dataset where the MQM scores generated by each annotator were the most similar as seen in **Table 21**.

Average MQM per LP		
LP	Average MQM	
	Annotator A	Annotator B
English-Japanese	91,7	72
English-Korean	40,7	84
English-Traditional Chinese	61,1	74,4
English-Simplified Chinese	83,4	77,1

Table 21. Average batch MQM per LP and annotator with the Unbabel Error Typology

This is due to the fact that, in most cases, while the annotators may identify the same errors they approach their annotation in completely different ways which, according to the guidelines for this typology, are not incorrect.

On the other hand, while agreement is high in the Japanese dataset, this LP presents a considerable gap in terms of MQM scores. This is mostly a result of the attribution of very different severities to the same errors, as will be explained in **examples 34** and **35** in the following section, which is not accounted for in terms of agreement with the Cohen Kappa coefficient.

The difference in MQM scores for the Korean dataset is compatible with its IAA results in the sense that, as mentioned before, *Annotator A* makes a considerable number of extra annotations in relation to *Annotator B* due to *Whitespace* errors. In addition, *Annotator B* is more lenient with severities and thus generates a much higher MQM with their annotations.

Finally, it is interesting to note that the Traditional Chinese dataset has relatively lower MQM scores compared to the other language pairs but that its corresponding IAA scores fall above the acceptable threshold, which is due to the fact that both annotators are more consistent in attributing higher severities to errors in relation to other language pairs.

5.1.2. Annotation Analysis

In relation to particles, which originated two new issue types in the East Asian Languages Annotation Module for the Unbabel Quality Framework as discussed in **Section 4.3.1.1.**, while using the Unbabel Error Typology in most cases the annotators for Japanese and Korean seemed to agree on annotating particles as prepositions, whether it was a case of omission or wrong usage.

Annotation of particles	
Japanese	Korean
EN (source) (16a) We have successfully cancelled the recurring payment with [PRODUCT]. JA (target) (16b) [PRODUCT]で定期支払いをキャンセルしました。 [Wrong Preposition, major]	EN (source) (17a) How many radios does the [PRODUCT] support? KO (target) (17b) [PRODUCT] 지원은 몇 개의 무선 신호 입니까? [Wrong Preposition, minor]

Table 22. Example of EN-JA annotation of particles with the Unbabel Error Typology

Although this is not grammatically correct and thus was a category addressed in the annotation module, apart from cases where one annotator completely ignores errors related to these words, the annotators in this dataset seem to be in consensus regarding the use of preposition related issue types for these errors.

The case of classifiers, however, is not the same as with particles, as the annotators seem to disagree regarding what issue type to use. In **examples 18 and 19**, *Annotator A* has annotated the omission of a classifier as an omitted determiner, while *Annotator B* used the *Other POS Omitted* issue type. As mentioned in **Section 4.3.1.1.** about the missing categories that were added to the East Asian Languages Annotation Module for the Unbabel Quality Framework, in the language guidelines for Simplified and Traditional Chinese classifiers are mentioned under the determiners section, which explains why *Annotator A* chose to use this issue type. However, seeing as this is not the correct category for these words and there is no other appropriate issue type, it is also justifiable that *Annotator B* considered it as *Other POS Omitted* instead.

Annotation of classifiers	
Annotator A	Annotator B
EN (source) (18a) Usually you can only see 1 wifi name on it. ZH-CN (target) (18b)通常您只能在上看到1无线网络名称。 [Omitted Determiner, minor]	EN (source) (19a) Usually you can only see 1 wifi name on it. ZH-CN (target) (19b) 通常您只能在上看到1无线网络名称。 [Other POS Omitted, major]

Table 23. Example of EN-ZH_CN annotation of classifiers with the Unbabel Error Typology

As explained previously in **Section 4.3.1.1.**, errors that involve transliteration have been annotated inconsistently in the past with varying issue types such as *Overly Literal* and *Lexical Selection*. In **example 20** from the annotations pertaining to the dataset annotated for this thesis, *Annotator B* for Korean has annotated what would be a *Transliteration* error as *Lexical Selection*. As *Annotator A* did not make any annotations regarding this word it is not possible to make a comparison between the issue types used by both, but it is still worth taking note of it in order to determine whether in the third dataset, annotated with the East Asian Languages Annotation Module for the Unbabel Quality Framework proposed on this thesis, *Annotator B* will correctly adopt the new issue type.

EN (source)

(20a) I have checked the details and I would like to inform you that as an one time exception gesture we can restore the deleted club.

KO (target)

(20b) 세부 사항을 확인했으며 일회성 예외 **체스처**로 삭제된 클럽을 복원할 수 있음을 알려드립니다.

[Lexical Selection, minor]

In addition to the errors related to the new categories that were added into the East Asian Languages Annotation Module for the Unbabel Quality Framework, the guidelines for the annotation module also attempt to address some of the issues that were registered on previous annotations and that were verified again on this dataset.

One of these issues were the cases of omission. *Omission* errors are frequently annotated inconsistently, particularly in the case of Japanese and both varieties of Chinese, due to the fact that these languages do not use whitespaces. In **examples 21** and **22** from the Japanese dataset, both annotators have annotated the same *Omission* error using different spans due to the fact that *Annotator A* considers the following words to be units that are separable from each other and *Annotator B* does not.

Annotation of omission	
Annotator A	Annotator B
EN (source) (21a) Their email services are still active. JA (target) (21b) メールサービスはまだ有効です。 [Other POS Omitted, major]	EN (source) (22a) Their email services are still active. JA (target) (22b) メールサービスはまだ有効です。 [Other POS Omitted, major]

Table 24. Example of different span selection in the annotation of the *Omission* type with the Unbabel Error Typology for JA

In the case of Simplified Chinese inconsistencies in *Omission* annotation were also found. In **examples 23** and **24**, more than one word was omitted before the highlighted units. While both annotators chose the *Other POS Omitted* issue type for this error, *Annotator B* used the additional issue type of *Omitted Conjunction*, thus making two annotations on the same error.

Annotation of omission	
Annotator A	Annotator B
EN (source) (23a) Just in case of chat disconnection, please don't hesitate to contact us again with your case number . ZH-CN (target) (23b) 如果聊天断开连接, 请随时与我们联系 个案号码 。 [Other POS Omitted, major]	EN (source) (24a) Just in case of chat disconnection, please don't hesitate to contact us again with your case number . ZH-CN (target) (24b) 如果聊天断开连接, 请随时与我们联系 个案号码 。 [Other POS Omitted, major] [Omitted Conjunction, major]

Table 25. Example of annotation of *Omission* with the Unbabel Error Typology for ZH-CN

Another issue that was explained at length in the guidelines for the annotation module due to previously registered errors was the annotation of verbs. In the past Japanese and Korean annotators in particular had the tendency to annotate what should be *Tense/Mood/Aspect* errors by marking only one component of the verb as *Addition* or use the appropriate issue type of *Tense/Mood/Aspect* but still annotating it on just one character that, if removed, would correct the error. In **example 25**, *Annotator B* of Japanese has used the method of annotating one component of the verb as *Addition* in order to indicate that if this character was removed the verb would be correct.

EN (source)

(25a) Please clear cache, change browser and try again later.

JA (target)

(25b) キャッシュをクリアして、ブラウザを変更して、後ほどもう一度お試しください。

[Addition, minor]

An additional point that was specifically addressed in the guidelines for the East Asian Languages Annotation Module for the Unbabel Quality Framework was the issue of punctuation. In Chinese and Japanese punctuation has to be full-width²⁸ and when it is not presented as such a *Punctuation* error should be annotated. In previously analyzed data annotators frequently ignored these types of errors and in the case of the annotations performed for this thesis this caused a high level of annotation disparity in the Simplified Chinese dataset. This dataset contains this kind of punctuation errors in almost every segment and while *Annotator A* annotated most of these errors, *Annotator B* ignored all of them, as seen in **Table 26**.



Annotation of punctuation in Simplified Chinese	
Annotator A	Annotator B
<p> 在我们仍在检查您的个案时,请继续等待,感谢您的耐心等待。</p> <p> 您好,顺便问一下,您有收据吗?</p>	<p>在我们仍在检查您的个案时,请继续等待,感谢您的耐心等待。</p> <p>您好,顺便问一下,您有收据吗?</p>

Table 26. Example of annotation of *Punctuation* in Simplified Chinese with the Unbabel Error Typology

Aside from the errors that were addressed in the guidelines for the East Asian Languages Annotation Module for the Unbabel Quality Framework, there are many other annotation disagreements between annotators that are predicted to be solved with this annotation module due to the fact that, in regard to many of the categories, it is much more simplified, as per

²⁸ Full-width punctuation is that which occupies the same space of a full character as opposed to half-width punctuation which occupies half the space of a character.

influence of the new version of the Unbabel Error Typology (v3) which was created after the integration with Lingo24. **Examples 26 to 29** show one instance for each of the four language pairs under discussion in which a mistranslation error was recognized by both annotators but ultimately annotated using different issue types, whether it was an issue type under *Mistranslation* or an external one.

Annotation of mistranslation errors			
Japanese	Korean	Traditional Chinese	Simplified Chinese
EN (source) (26a) No worries. JA (target) (26b) <u>ご心配には及び びません。</u> [A: Overly Literal] [B: Lexical Selection]	EN (source) (27a) Korean is not working earlier. KO (target) (27b) <u>중국어가 이전에는 작동하지 않습니다.</u> [A: Lexical Selection] [B: MT Hallucination]	EN (source) (28a) Hi there, Nice to meet you! ZH-TW (target) (28b) <u>嗨, 您好, 尼斯 去了!</u> [A: Lexical Selection] [B: MT Hallucination]	EN (source) (29a) In our website it's not yet available. ZH-CN (target) (29b) <u>在我们的网站 上它还没有可用。</u> [A: Lexical Selection] [B: POS]

Table 27. Examples of annotation of *Mistranslation* with the Unbabel Error Typology across all LPs

Similarly, *Register* errors were also annotated inconsistently due to their division into *Lexical* and *Grammatical Register* which was not clear to some annotators, as seen in **examples 30 to 33**.

Annotation of register	
Annotator A	Annotator B
EN (source) (30a) Hey [NAME]! KO (target) (30b) 이봐 [NAME]! [Lexical Register, major]	EN (source) (31a) Hey [NAME]! KO (target) (31b) 이봐 [NAME]! [Grammatical Register, major]
EN (source) (32a) Hi there, [NAME] . ZH-TW (target) (32b) 你好, [NAME]。 [Grammatical Register, major]	EN (source) (33a) Hi there, [NAME] . ZH-TW (target) (33b) 你好, [NAME]。 [Lexical Register, major]

Table 28. Examples of EN-KO and EN-ZH_TW annotation of *Register* with the Unlabel Error Typology

Finally, it is also important to mention that in many cases where the annotators agreed upon the issue type that should be used, there is disagreement in terms of severities.

Choice of severities	
Annotator A	Annotator B
EN (source) (34a) If you are still unable to make payment, please contact your card issuer. JA (target) (34b) それでもお支払いができない場合は、カード 発行者 にお問い合わせください。 [Lexical Selection, major]	EN (source) (35a) If you are still unable to make payment, please contact your card issuer. JA (target) (35b) それでもお支払いができない場合は、カード 発行者 にお問い合わせください。 [Lexical Selection, minor]
EN (source) (36a) For you to remove the current one, you have to contact the firmware provider. ZH-CN (target) (36b) 为了删除当前的 one , 您必须与固件提供商联系。 [Untranslated, critical]	EN (source) (37a) For you to remove the current one, you have to contact the firmware provider. ZH-CN (target) (37b) 为了删除当前的 one , 您必须与固件提供商联系。 [Untranslated, major]

Table 29. Examples of mismatch of severities with the Unbabel Error Typology in Japanese and Simplified Chinese

Although in some cases, such as the double annotation instance on one *Omission* error mentioned on **example 24**, the inconsistencies in the annotations are related to poor interpretation of the annotation guidelines, there are many instances where the result of these annotations prove the need for a new typology and guidelines as the current ones are unclear at times.

5.1.3. Annotators' feedback

Previous to the annotations that were performed for the experiments on this thesis, there had been annotation surveys for several language pairs within Unbabel in order to gain knowledge on which issue types annotators felt were not applicable to their respective language

pairs and which additional issue types they believed were missing in the typology. Among other languages these surveys gathered the input of the annotators for the translation directions of English to Japanese and English to Simplified Chinese. **Table 30** represents which issue types were pointed out as not applicable for both language pairs, while **Table 31** demonstrates the issue types that were proposed for each LP by the annotators.

Annotators' Feedback		
Not applicable issue types	Japanese	Simplified Chinese
Ambiguous Translation	✓	
Overly Literal	✓	
False Friend	✓	✓
Source/Target Disagreement	✓	✓
Wrong Paronym	✓	
POS	✓	
Capitalization	✓	
Diacritics	✓	✓
Hyphenation	✓	✓
Wrong Language Variety	✓	
Hypernym/Hyponym	✓	
Synonym	✓	
Mistranslated Term	✓	
Term Wrongly Applied		✓

Tense/Mood/Aspect		✓
Omitted Auxiliary Verb	✓	
Omitted Determiner	✓	
Agreement	✓	

Table 30. Annotators' feedback on not applicable issue types in the Unlabel Error Typology

Annotators' feedback		
Proposed issue types	Japanese	Simplified Chinese
Inappropriate	✓	
Omitted Aspect Marker		✓
Omitted Argument		✓
Omitted Adjunct		✓
Omitted Particle		✓
Omitted Classifier		✓
Wrong Classifier		✓

Table 31. Annotators' feedback on missing issue types in the Unlabel Error Typology

When analyzing this table it becomes clear that the annotators of both languages consider many of the specific issue types under *Mistranslation* as unnecessary and confusing, as some of these were also pointed out to be unclearly defined in the annotation guidelines for this typology. In addition, the issue types contained under *Spelling* were also considered as useless by the annotators of both language pairs, either because the usage guidelines are unclear or because the issue types do not apply, in the case of Chinese. Similarly, issue types such as *Hyphenation*, *Diacritics* and *Capitalization* were also identified as unnecessary due to the characteristics of the

languages at hand. Furthermore, the Simplified Chinese annotators also identified *Tense/Mood/Aspect* as an unnecessary issue type, as verbs in Chinese are not inflected and *Tense/Mood/Aspect* is expressed through the usage of particles, adverbs and auxiliary verbs. Finally, it is also visible that *Terminology* issue types have been marked as unnecessary. While this is not due to the linguistic characteristics of any of these translation directions, as terminology is a part of both, it is likely that the existence of three different issue types under *Terminology* is confusing for annotators who end up considering at least one of them as superfluous.

In relation to the issue types the annotators suggested should be added, the Chinese annotators requested further distinction of categories within *Omission* and one additional issue type under *Function Words* for wrong classifiers while the Japanese annotators only suggested one new issue type named “Inappropriate” which, in the proposed definition, was very similar to the already existing *Overly Literal* issue type.

However, it should be noted that during the round of annotations for this thesis that were performed using the Unbabel Error Typology, one of the Japanese annotators was the only to leave feedback regarding this typology, in which they asked precisely to report missing issue types, using *Missing Classifier* as an example.

When analyzing the results of the annotations performed with this typology, it was necessary to keep in mind two important factors that could have opposing effects on the results.

On one hand, this typology was the most extensive of all three under analysis, with a total of 47 selectable issue types, and it is also the typology that was considered to be in need of an adaptation that was more suitable for the annotation of East Asian languages.

On the other hand, however, it was also the typology the annotators were the most familiar with.

This means that while there was already concrete feedback on this typology and the changes that it needed in the eyes of the annotators, it was expected that due to familiarity with the typology the IAA results in particular would not be extremely low, which was verified mostly with Japanese but did not verify with the Simplified Chinese annotations in particular, which

might be due to the fact that, as seen in **Section 5.1.3.**, it is the language pair for which the annotators made more remarks in relation to missing categories.

5.2. Ye & Toral’s (2020) Proposal

The second typology used for the annotation experiments on this thesis was the MQM-compliant error taxonomy proposed by Ye and Toral (2020) specifically for the translation direction of English to Chinese. Despite being conceived for annotation of the English-Chinese LP, in this phase of annotations the typology was used to annotate all four LPs which have already been mentioned.

5.2.1. Inter-annotator Agreement

Table 32 represents the average IAA scores obtained for each language pair with the typology proposed by Ye and Toral and in comparison with the values obtained with the Unbabel Error Typology.

Average batch IAA per Language Pair (LP)				
LP	Average batch IAA		% of jobs above 0.4 threshold	
	Ye and Toral	Unbabel	Ye and Toral	Unbabel
English-Japanese	0.526	0.628	58,3%	54,5%
English-Korean	0.454	0.366	8,3%	50%
English-Traditional Chinese	0.352	0.464	28,6%	60%
English-Simplified Chinese	0.194	0.192	5%	17,6%

Table 32. Average batch IAA per LP with Unbabel’s and Ye and Toral’s typologies

As seen in **Table 32**, the IAA scores for the English to Japanese and English to Traditional Chinese datasets were lower when annotated using the typology proposed by Ye and Toral in comparison with the Unbabel Error Typology and in the case of Traditional Chinese this reduction meant that the score is no longer above the threshold established as acceptable. In the case of the English to Korean and English to Simplified Chinese datasets the IAA scores were improved and, while the increase for Simplified Chinese was minimal, for Korean it meant that the average score rose above the 0.4 threshold.

However, for all LPs except English to Japanese the percentage of jobs with IAA scores above 0.4 decreased considerably. Furthermore, by comparing the IAA for each individual job between the two typologies it was possible to conclude that the number of cases where the IAA scores increased or decreased is very similar.

One of the characteristics that separates this typology from the Unbabel Error Typology most notably and that had a great impact on annotation was the fact that it does not contain issue types related to *Whitespace* and *Register*. As will be discussed in more detail in the **Section 5.2.3.**, almost all annotators pointed out that these issue types are very necessary when annotating the languages at hand and that annotations suffered from their non-existence.

This is particularly evident in the case of one of the Korean annotators, as they reported it was impossible to annotate certain jobs without these issue types and with the lack of appropriate *Mistranslation* issue types. In light of this, the annotator did not annotate several jobs, making it so that it was not possible to measure IAA for them. As such, the average IAA score calculations were affected by this since many of the jobs were not taken into account.

The lack of the issue types of *Whitespace* and *Register* also affected the agreement for Japanese. While the two annotators continued to have a fair level of agreement while using the typology proposed by Ye and Toral, the average score ended up being lower due to the disappearance of a large number of *Whitespace* annotations which were consensual and made the agreement be higher when annotating with the Unbabel Error Typology.

The case of Traditional Chinese is more complex than with the other languages due to a few factors. Firstly, due to constraints beyond our control it was necessary to switch *Annotator B* after the annotation process with the Unbabel Error Typology. This means that *Annotator B* that annotated with the Unbabel Error Typology is not the same *Annotator B* that annotated with the

Ye and Toral's typology and, thus, the annotations pertaining to both typologies have some natural differences, the most significant one being that, in general, the new *Annotator B* annotates more than the previous one. In relation to the agreement between *Annotator A* and *B* while using this typology, the most relevant reason why it decreased is the fact that, as shown in **examples 38 to 41**, *Annotator B* applied the *Overly Literal* issue type to all kinds of errors that did not have a directly corresponding issue type, while *Annotator A* used other additional issue types, such as *Unintelligible*.

Issue type disagreement	
Annotator A	Annotator B
EN (source) (38a) Nice to meet you. ZH-TW (target) (38b) 很高興見到您。 [Overly Literal, major]	EN (source) (39a) Nice to meet you. ZH-TW (target) (53b) 很高興見到您。 [Overly Literal, major]
EN (source) (40a) Your paid plan will resume and you will be charged on March 28. ZH-TW (target) (40b) 您的付費計劃將恢復，並且您將在 283 月向您收費。 [Unintelligible, critical]	EN (source) (41a) Your paid plan will resume and you will be charged on March 28. ZH-TW (target) (41b) 您的付費計劃將恢復，並且您將在 283 月向您收費。 [Overly Literal, major]

Table 33. Examples of issue type disagreement with Ye and Toral's typology in Traditional Chinese

Finally, in similarity with what was verified with the Unbabel Error Typology, in the case of Simplified Chinese the low IAA is in part due to the numerous occurrences of punctuation errors that *Annotator A* identifies and *Annotator B* overlooks. In addition, a level of lack of understanding of the annotators in relation to the organization of grammar errors in the typology proposed by Ye and Toral, which will be discussed in the following section, also affects IAA

negatively for this language pair as the annotators make use of the *Function Words* issue types a great number of times, yet not in consensus. However, this was also one of LPs where the fact that *Omission* is one single issue type with no daughter tags had the most positive results.

Table 34 represents the average MQM score corresponding to each annotator and LP while annotating with this typology.

Average MQM per LP				
LP	Average MQM			
	Ye and Toral (2020)		Unbabel Error Typology	
	Annotator A	Annotator B	Annotator A	Annotator B
English-Japanese	92	86	91,7	72
English-Korean	62,7	79,2	40,7	84
English-Traditional Chinese	73,75	79,8	61,1	74,4
English-Simplified Chinese	84,4	81,7	83,4	77,1

Table 34. Average batch MQM per LP and annotator with Ye and Toral’s typology

In all cases except that of *Annotator B* for Korean the MQM scores increased in relation to the annotations made with the Unbabel Error Typology. This is due to the fact that, as mentioned before, this typology does not contain issue types for errors that are very common in these datasets, such as *Whitespace* and *Register*. As such, even though the quality of the datasets is the same as before their quality appears to be higher due to the fact that some errors could not be annotated, which is one of the consequences of using an inadequate typology.

5.2.2. Annotation Analysis

As the error typology proposed by Ye and Toral (2020) was conceived specifically for annotation in the translation direction of English to Chinese, it already contains issue types that are compliant with some of the characteristics of the Chinese language, as is the case of particles and classifiers.

As shown in **Section 3.4.3.**, one of the biggest differences between this typology and the Unbabel Error Typology is the way the annotation of *Function Words* is organized. In the case of the typology proposed by Ye and Toral (2020), aside from the issue types under *Extraneous*, which are meant to be used in cases where Chinese expressions appear too westernized in the target due to the overuse of function words, there are two additional issue types of *Incorrect* and *Missing*. As such, the way to correctly annotate using this typology would be to tag any omitted function word as *Missing* and function words that are being wrongly used as *Incorrect*.

Furthermore, under *Grammar* there is also the issue type of *Classifier* which, as per the guidelines that were created for the annotators, should be used to annotate the incorrect use of classifiers.

However the division of the categories under *Grammar*, including those under *Function Words*, seems to confuse the annotators due to a few different factors. Firstly, the fact that there is a *Missing* issue type exclusive for *Function Words* in addition to the general *Omission* issue type that also exists in the typology caused some disagreement. As seen in **examples 42** and **43**, where a function word was omitted in the Japanese target, while *Annotator A* used the *Missing Function Word* issue type, *Annotator B* used *Omission* to annotate it.

Annotation of omitted function words	
Annotator A	Annotator B
EN (source) (42a) This is the transaction ID ALPHANUMERICID-0. JA (target) (42b) 取引 ID [Ø]ALPHANUMERICID-0です。 [Missing Function Word, minor]	EN (source) (43a) This is the transaction ID ALPHANUMERICID-0. JA (target) (43b) 取引 ID [Ø]ALPHANUMERICID-0です。 [Omission, minor]

Table 35. Example of EN-JA annotation of omitted function words with Ye and Toral’s typology

However, there were also instances where *Annotator B* made use of the *Missing* and *Incorrect Function Word* issue types where they had previously annotated the same errors in relation to prepositions while using the Unbabel Error Typology, as seen in the following example.

Annotation of function words	
Unbabel Error Typology	Typology proposed by Ye and Toral
EN (source) (16a) We have successfully cancelled the recurring payment with [PRODUCT]. JA (target) (16b) [PRODUCT]で定期支払いをキャンセルしました。 [Wrong Preposition, major]	EN (source) (44a) We have successfully cancelled the recurring payment with [PRODUCT]. JA (target) (44b) [PRODUCT]で定期支払いをキャンセルしました。 [Incorrect Function Word, major]

Table 36. Example of EN-JA annotation of function words with Ye and Toral’s typology

This implies that the understanding the annotators have over this typology is flawed, either due to the fact that it is an unfamiliar taxonomy or that the organization of the above-mentioned issue types is not intuitive.

In fact, the case of classifiers is similar in the sense that the annotators disagree on which issue type should be used for each error, particularly in the cases of *Addition* and *Omission*. In **examples 45** and **46**, where *Annotator B* for Simplified Chinese used the *Classifier* issue type to identify the omission of a classifier in the target, *Annotator A* used the *Missing Function Word* issue type.

Annotation of classifiers	
Annotator A	Annotator B
EN (source) (45a) Usually you can only see 1 wifi name on it. ZH-CN (target) (45b)通常您只能在上看到1无线网络名称。 [Missing Function Word, minor]	EN (source) (46a) Usually you can only see 1 wifi name on it. ZH-CN (target) (46b)通常您只能在上看到1无线网络名称。 [Classifier, minor]

Table 37. Example of annotation of EN-ZH_CN classifiers with Ye and Toral’s typology

As per the guidelines the annotators were provided with, the *Classifier* issue type is defined as such:

Classifier

Issues related to the incorrect use of classifiers. Classifiers are special linguistic units located behind a number, demonstrative or certain quantifiers. These classifiers do not have a counterpart in English, which might give rise to translation problems.

Ex:

(EN) Click the New Account pop-up menu, then choose a type of user.

(ZH-CN) 点击新账户弹出菜单，然后选择一个^{CLASSIFIER}类型的用户。

(ZH-CN) 点击新账户弹出菜单，然后选择一种^{CLASSIFIER}类型的用户。

Table 38. Definition of *Classifier* errors in Ye and Toral’s typology

The fact that the annotators chose different issue types for the errors, shown in **examples 45** and **46**, demonstrates that *Annotator B* understood the omission of classifiers as being a case of their “incorrect use”, while *Annotator A* considered incorrect use as having the wrong classifier on the text as presented in the example under the definition.

Although the *Grammar* category on this typology was the source of several points of disagreement between annotators, *Mistranslation* was the category that proved to be more problematic. This is due to the fact that *Overly-literal* and *Entity* are only two issue types under *Mistranslation* in this typology and that *Mistranslation* as a parent issue type was not selectable. As such, in most cases of *Mistranslation* errors, in comparison to how they were annotated using the Unbabel Error Typology, annotations were much more inconsistent. As seen in **examples 47** and **48**, some errors that were unanimously annotated as *Lexical Selection* with the Unbabel Error Typology were separated into two different issue types. In this case, *Unintelligible* was chosen by *Annotator A* even though it is not under *Mistranslation*.

Annotation of mistranslation	
Unbabel Error Typology	Typology proposed by Ye and Toral
EN (source) (47a) Upon checking, the card went missing because it was disabled. JA (target) (47b) 確認いたしましたところ、カードは無効になっておりますので、 足りなくな っております。 [Annotator A: Lexical Selection] [Annotator B: Lexical Selection]	EN (source) (48a) Upon checking, the card went missing because it was disabled. JA (target) (48b) 確認いたしましたところ、カードは無効になっておりますので、 足りなくな っております。 [Annotator A: Unintelligible] [Annotator B: Overly Literal]

Table 39. Example of EN-JA annotation of *Mistranslation* with Ye and Toral’s typology

Similarly, a considerable amount of errors that were annotated differently between annotators with the Unbabel Error Typology continued to be separated into different issue types by the annotators while using Ye and Toral’s proposed typology.

Annotation using different issue types	
Unbabel Error Typology	Typology proposed by Ye and Toral
EN (source) (27a) Korean is not working earlier. KO (target) (27b) <u>중국어</u> 가 이전에는 작동하지 않습니다. [A: Lexical Selection] [B: MT Hallucination]	EN (source) (49a) Korean is not working earlier. KO (target) (49b) <u>중국어</u> 가 이전에는 작동하지 않습니다. [A: Unintelligible] [B: Entity]
EN (source) (28a) Hi there, Nice to meet you! ZH-TW (target) (28b) 嗨, 您好, <u>尼斯去了!</u> [A: Lexical Selection] [B: MT Hallucination]	EN (source) (50a) Hi there, Nice to meet you! ZH-TW (target) (50b) 嗨, 您好, <u>尼斯去了!</u> [A: Unintelligible] [B: Overly Literal]

Table 40. Examples of EN-KO and EN-ZH_TW annotation using different issue types with Ye and Toral’s typology

Furthermore, as there are no issue types concerning *Terminology* errors²⁹ in this typology, in the cases they occurred in the text the annotators also disagreed in their approach. As seen in **examples 51** and **52**, while *Annotator A* for Simplified Chinese annotated a terminology error using the *Unintelligible* issue type, *Annotator B* considered that the lack of terminology related issue types made it impossible to annotate them and, as such, left the segment without annotations.

²⁹ *Terminology* errors are those that refer to the glossary terms. They appear highlighted in blue on the Annotation Tool so that the annotators can recognize them.

Annotation of terminology	
Annotator A	Annotator B
<p>EN (source)</p> <p>(51a) We love to help you remove it, but we don't have the resources to support this [PRODUCT] since the router is an open source.</p> <p>ZH-CN (target)</p> <p>(51b) 我们很乐意帮助您移除它, 但由于路由器是开放的 zh-CN , 因此我们没有支持此 [PRODUCT] 的资源。</p> <p>[Unintelligible, major]</p>	<p>EN (source)</p> <p>(52a) We love to help you remove it, but we don't have the resources to support this [PRODUCT] since the router is an open source.</p> <p>ZH-CN (target)</p> <p>(52b) 我们很乐意帮助您移除它, 但由于路由器是开放的 zh-CN , 因此我们没有支持此 [PRODUCT] 的资源。</p> <p>[no annotations]</p>

Table 41. Example of EN-ZH_CN annotation of *Terminology* with Ye and Toral’s typology

However, the fact that this typology is much smaller than the one used at Unbabel also reduces the probability of disagreement, as there are less issue types to choose from and the granularity of the typology is also reduced. As such, some annotations that were previously different between annotators were unified while using the typology proposed by Ye and Toral, as seen on **examples 26 and 53 to 55**.

Unification of Issue Types	
Unbabel Error Typology	Ye and Toral's typology
EN (source) (26a) No worries. JA (target) (26b) <u>ご心配には及びません。</u> [A: Overly Literal] [B: Lexical Selection]	EN (source) (53a) No worries. JA (target) (53b) <u>ご心配には及びません。</u> [A: Overly Literal] [B: Overly Literal]
Source (54a) <u>支払いは競合の影響を受けませんのでご安心ください。</u> (54b) <u>支払い橋合同教室を慈善団体楽天会</u> [A: Shouldn't have been translated] [B: MT Hallucination]	Source (55a) <u>支払いは競合の影響を受けませんのでご安心ください。</u> (55b) <u>支払い橋合同教室を慈善団体楽天会</u> [A: Unintelligible] [B: Unintelligible]

Table 42. Examples of unification of issue types with Ye and Toral's typology in Japanese

Finally, it is important to mention that this typology does not contain issue types for *Whitespace* and *Register* errors. As such, the total number of annotations is very reduced in comparison to the ones made using the Unbabel Error Typology, as these were very common errors in the dataset. As will be discussed in the following sections, this had an impact that was pointed out by almost all the annotators and was reflected on IAA.

5.2.3. Annotators' Feedback

The feedback the annotators gave in relation to this typology was always related to the missing issue types mentioned in the previous sections that the annotators were accustomed to use when annotating with the Unbabel Error Typology. The annotators for all LPs except English to Japanese asked questions about or commented on the lack of issue types, particularly in relation to subtypes of *Mistranslation* in addition to *Whitespace* and *Register* errors. **Table 43**

identifies the issue types that were noted as missing and necessary to properly annotate the jobs per LP.

Annotators' Feedback on Missing Issue Types				
Missing Issue Types	Japanese	Korean	Traditional Chinese	Simplified Chinese
Mistranslation issue types (such as Lexical Selection)		✓		✓
Wrong Number				✓
Register		✓	✓	✓
Terminology				✓
Whitespace		✓		✓
Wrong Language Variety			✓	

Table 43. Annotators' feedback on missing issue types in Ye and Toral's typology

As seen in **Table 43**, the issue type that all annotators except those for Japanese pointed out as missing was *Register*. However, while the English to Japanese annotators did not leave any comments and this continued to be the LP with the highest IAA scores, it was possible to conclude that they also struggled with *Register* errors due to the fact that there had been some *Register* annotations done using the Unbabel Error Typology that were not possible to annotate using the typology proposed by Ye and Toral.

At the same time, the lack of options under *Mistranslation* and an issue type for *Whitespace* were also pointed out by more than one annotator, including the Simplified Chinese annotators which were working with the translation direction the typology was conceived for. This is due to the fact that, even though whitespaces are not used when writing in Chinese, machine translations may generate misplaced whitespaces that have to be annotated and make this issue type necessary.

As for *Wrong Language Variety* it was only pointed out by the Traditional Chinese annotators, which was expected due to the fact that these errors are not relevant in the case of Japanese and Korean.

This feedback's reflection on the annotations was diverse. While the annotators for Japanese and Traditional Chinese used the typology as it was and only refrained from annotating errors such as *Whitespace* and *Register*, as it was impossible to do so, annotators B for Korean and Simplified Chinese proceeded differently.

In the case of Simplified Chinese, *Annotator B* used the Comment option on the Annotation Tool to list all the errors they believed could not be annotated due to the lack of appropriate issue types. Most of the issues listed were *Mistranslation* errors.

On the other hand, the case of Korean was more severe in the sense that *Annotator B* considered the lack of issue types was too great and, as such, completely refrained from annotating a number of jobs.

While this typology introduced issue types that were very relevant for annotation of the languages under discussion, like *Classifier* and *Particle* under *Function Words*, the issue types that it lacks affected its usability from the point of view of the annotators. While issue types such as *Terminology* are not necessary in all contexts, in the case of the content types translated by Unbabel it is an essential category. At the same time, issue types like *Whitespace* and *Register* are necessary for annotation of these languages in all contexts, particularly in the case of MT, and should be a part of the typology.

The case of *Mistranslation* is more subjective, as it could be argued that the typology proposed by Ye and Toral meant for the *Mistranslation* issue type itself to be selectable even though it is further divided into two other issue types. However, due to the limitations of the Annotation Tool it was not possible to define it as selectable. As such, it is more important to look at the results from the point of view of *Grammar*, which is the error category where the most relevant differences between Western languages and the East Asian languages under analysis can be found. In the typology proposed by Ye and Toral (2020) this is also the category that includes the issue types of *Classifier* and those under *Function Words* that were included specifically for the annotation of Chinese.

Through the annotation results it was possible to conclude that the annotators made prolific use of these issue types which were unfamiliar to them and did not have any questions about them, which shows that they understood their necessity as confirmed in the feedback provided in relation to the Unbabel Error Typology. However, the organization of these issue types seemed to raise disagreement, as seen in **examples 42-43** and **45-46**. This means that while it is important to have these issue types in a typology for annotation of East Asian languages, it is also necessary to have a more clear arrangement that makes their use more intuitive.

5.3. East Asian Languages Annotation Module for the Unbabel Quality Framework

Finally, the third and last round of annotations was performed using the East Asian Languages Annotation Module for the Unbabel Quality Framework whose creation was the main focus of this internship.

5.3.1. Inter-annotator Agreement

Table 44 represents the average IAA scores corresponding to the East Asian Languages Annotation Module for the Unbabel Quality Framework in comparison with the previously analyzed scores obtained with Unbabel's and Ye and Toral's typologies.

Average batch IAA per Language Pair (LP)						
LP	Average batch IAA			% of jobs above 0.4 threshold		
	East Asian Languages Annotation Module	Ye and Toral	Unbabel	East Asian Languages Annotation Module	Ye and Toral	Unbabel
English-Japanese	0.526	0.526	0.628	66,7%	58,3%	54,5%/ /
English-Korean	0.355	0.454	0.366	31,3%	8,3%	50%
English-Traditional Chinese	0.520	0.352	0.464	68,8%	28,6%	60%
English-Simplified Chinese	0.189	0.194	0.192	5%	5%	17,6%

Table 44. Average batch IAA per LP with the Unbabel Error Typology, the typology proposed by Ye and Toral and the East Asian Languages Annotation Module for the Unbabel Quality Framework

From the results shown in the **Table 44** it was possible to conclude that even though the case of Traditional Chinese was positive, as both the average IAA and the percentage of jobs with acceptable IAA increased, this annotation module still needs to be improved in order to allow the same to happen for the other language pairs. It is important to note, however, that in the case of Japanese, even though the average IAA score decreased in relation to that corresponding to the Unbabel Error Typology, it still remains above the 0.4 threshold and the percentage of individual jobs above this threshold actually increased.

While it was expected that this annotation module would produce similar IAA results between Traditional and Simplified Chinese due to their similarity, this was not the case. In fact, Simplified Chinese was the LP for which all results were persistently low across all typologies, with the average IAA decreasing even further when using the East Asian Languages Annotation Module for the Unbabel Quality Framework. Thus, it was important to look at annotations individually to determine where the biggest points of improvement, in the case of Traditional

Chinese, and major issues, in the case of Simplified Chinese, lie in order to improve the points where the annotation module performed more poorly.

Although some confusion still remains in terms of the categories within *Mistranslation*, Traditional Chinese was one of the LPs that benefited the most from the reduction of issue types under it as well as the simplification of *Register* to one single issue type.

On the other hand, the biggest reason why the Simplified Chinese IAA scores were the lowest was the fact that the jobs for this LP contain an extremely high number of *Punctuation* errors which were initially not annotated at all by one of the annotators while using Unbabel's and Ye and Toral's typologies. With the East Asian Languages Annotation Module for the Unbabel Quality Framework, while the errors were identified by both annotators, they still annotated them differently due to poor interpretation of the guidelines, as will be shown in the following section in **Table 57**. In addition, the annotators disagree in relation to the categories within *Omission* and *Mistranslation* as they sometimes use different issue types for the same errors.

The case of Japanese is interesting in the sense that while the average IAA score decreased, the percentage of jobs which scored above the 0.4 threshold was higher. Upon analyzing and comparing annotations, it could be concluded that *Annotator A* was more consistent with their annotations across the different typologies while *Annotator B* annotated some errors in different ways depending on the typology, as seen in **examples 56** and **57**.

Agreement on register and terminology	
Unbabel Error Typology	East Asian Languages Annotation Module
EN (source) (56a) We will close this chat for now and please let us know if you have further questions.	EN (source) (57a) We will close this chat for now and please let us know if you have further questions.
JA (target) (56b) こちらのチャットを閉じますので、質問がございましたらお知らせください。 [Lexical Register, major] [Other POS Omitted, major]	JA (target) (57b) こちらのチャットを閉じますので、質問がございましたらお知らせください。 [Register, major]

Table 45. Example of EN-JA annotation of *Register* and *Terminology* with the East Asian Languages Annotation Module for the Unbabel Quality Framework

Similarly to what happened with Traditional Chinese, however, the reduction of issue types under *Mistranslation* helped improve IAA scores for the English-Japanese annotations as well. In addition, this was also the LP that benefited the most from what was pointed out in the guidelines in relation to span selection.

In relation to the IAA scores for Korean, the approach to the annotation of whitespaces was what influenced IAA scores negatively the most. The Korean dataset is heavily polluted with *Whitespace* errors which *Annotator B* mostly overlooked, likely due to the fact that they were extremely frequent. This became even more evident when annotating with the East Asian Languages Annotation Module for the Unbabel Quality Framework, as *Annotator A* continued to annotate every *Whitespace* error in the batch while *Annotator B* annotated even less than before. The weight of this difference is visible in the MQM scores as well, as *Annotator B* consistently generated higher MQM scores in the jobs more polluted with *Whitespace* errors, which also resulted in English-Korean being the only LP where *Annotator A* is the one who presents the lower MQM score out of the two, as seen in **Table 46**.

Average MQM per LP						
LP	Average MQM					
	East Asian Languages Annotation Module		Ye and Toral (2020)		Unbabel Error Typology	
	Annotator A	Annotator B	Annotator A	Annotator B	Annotator A	Annotator B
English-Japanese	91,7	77	92	86	91,7	72
English-Korean	42,6	60,5	62,7	79,2	40,7	84
English-Traditional Chinese	65	60,8	73,75	79,8	61,1	74,4
English-Simplified Chinese	84,5	75,2	84,4	81,7	83,4	77,1

Table 46. Average batch MQM per LP and annotator with the East Asian Languages Annotation Module for the Unbabel Quality Framework

It is important to note that in Ye and Toral (2020) there is no mention of severities and, as such, for the purpose of the experiments on this thesis the same three severity levels which are a part of the Unbabel Error Typology and the East Asian Languages Annotation Module were also used for this typology with the same weights. Aside from allowing a faithful comparison of MQM scores, the attribution of severities is essential from the point of view of typology benchmarking and transparency, so it was important to include it for all typologies. However no mention of severities was included in the guidelines corresponding to the mentioned typology and their selection was left entirely up to the annotators' judgment.

Finally, it is important to point out that it was expected that the typology proposed by Ye and Toral (2020) would have the highest IAA scores out of all typologies, due to the fact that it has significantly less categories than both the Unbabel Error Typology and the East Asian Languages Annotation Module, as can be seen in the table on **Annex A**. Despite this, as can be seen on **Table 44**, the IAA scores corresponding to Ye and Toral's typology were either lower

than those corresponding to the other typologies or not too different, depending on the language pair. Furthermore, for all language pairs except English-Simplified Chinese this typology had fewer jobs falling above the acceptable threshold than the East Asian Languages Annotation Module. In light of this, it can be concluded that in terms of typologies designed for a specific set of languages the East Asian Languages Annotation Module performed better.

5.3.2. Annotation Analysis

As discussed in **Section 4.3.2.**, the East Asian Languages Annotation Module for the Unbabel Quality Framework was created by taking the main structure from the newest version of the Unbabel Error Typology (v3) which is currently in use at Unbabel and adding issue types that were relevant for the language pairs under discussion and, similarly, removing issue types that were not applicable in order to reduce unnecessary noise.

Firstly, it is important to verify if, and how, the annotators used the new categories.

In relation to the two issue types about particles, there was a clear transition of errors previously annotated as prepositions to these issue types, as seen in **examples 58-61**.

Annotation of particles	
Unbabel Error Typology	East Asian Languages Annotation Module
EN (source) (58a) We have successfully cancelled the recurring payment with [PRODUCT]. JA (target) (58b) [PRODUCT]で定期支払いをキャンセルしました。 [Wrong Preposition, major]	EN (source) (59a) We have successfully cancelled the recurring payment with [PRODUCT]. JA (target) (59b) [PRODUCT]で定期支払いをキャンセルしました。 [Wrong Particle, major]
EN (source) (60a) What is the [PRODUCT] ALPHANUMERICID-0 Dual-Band Smart Wi-Fi Wireless Router? KO (target) (60b) [PRODUCT] ALPHANUMERICID-0 듀얼 밴드 Smart Wi-Fi 무선 공유기은 무엇입니까? [Wrong Preposition, minor]	EN (source) (61a) What is the [PRODUCT] ALPHANUMERICID-0 Dual-Band Smart Wi-Fi Wireless Router? KO (target) (61b) [PRODUCT] ALPHANUMERICID-0 듀얼 밴드 Smart Wi-Fi 무선 공유기은 무엇입니까? [Wrong Particle, minor]

Table 47. Examples of EN-JA and EN-KO annotation of particles with the East Asian Languages Annotation Module for the Unbabel Quality Framework

The classifier issue types were also successfully applied. As seen in **examples 62 to 64**, where classifiers were annotated previously using two different issue types when using both the Unbabel Error Typology and the taxonomy proposed by Ye and Toral, with the East Asian Languages Annotation Module for the Unbabel Quality Framework the annotators are in agreement.

Annotation of classifiers	
Unbabel Error Typology	Ye and Toral's Typology
EN (source) (62a) Usually you can only see 1 wifi name on it. ZH-CN (target) (62b) 通常您只能在上看到1无线网络名称。 [A: Omitted Determiner, minor] [B: Other POS Omitted, major]	EN (source) (63a) Usually you can only see 1 wifi name on it. ZH-CN (target) (63b) 通常您只能在上看到1无线网络名称。 [A: Missing Function Word, minor] [B: Classifier, minor]
East Asian Languages Annotation Module	
EN (source) (64a) Usually you can only see 1 wifi name on it. ZH-CN (target) (64b) 通常您只能在上看到1无线网络名称。 [A: Omitted Classifier, minor] [B: Omitted Classifier, major]	

Table 48. Example of EN-ZH_CN annotation of classifiers with the East Asian Languages Annotation Module for the Unbabel Quality Framework

The *Transliteration* issue type, however, appears to not have been understood equally by all annotators. In the case of Korean, while annotating with the Unbabel Error Typology *Annotator B* had identified an error that would classify as *Transliteration* with the East Asian Languages Annotation Module for the Unbabel Quality Framework. However, in the last round of annotations the annotator not only did not apply any error tag to that error but also used the *Transliteration* issue type on an error to which it did not apply, as seen in **example 65** and **66**. On the other hand, *Annotator B* for Traditional Chinese correctly applied the issue type to an error, although it was not in agreement with *Annotator A* who used *Overly Literal* for the same error, as seen in **example 67** and **68**.

Annotation of transliteration	
Annotator A	Annotator B
EN (source) (65a) Sounds good? KO (target) (65b) 좋은 소리? [Overly Literal, major]	EN (source) (66a) Sounds good? KO (target) (66b) 좋은 소리? [Transliteration, major]
EN (source) (67a) Hi there [NAME] , Nice to meet you! ZH-TW (target) (67b) 嗨, 您好 John, 尼斯去了! [Lexical Selection, critical]	EN (source) (68a) Hi there [NAME] , Nice to meet you! ZH-TW (target) (68b) 嗨, 您好 John, 尼斯去了! [Transliteration, major]

Table 49. Example of EN-KO and EN-ZH_TW annotation of *Transliteration* with the East Asian Languages Annotation Module for the Unbabel Quality Framework

In fact, one of the categories in which this annotation module did not perform as well as expected was *Mistranslation*. Although it is a category that is much more reduced in comparison to the one in Unbabel Error Typology, while the annotators came to agree on some annotations as seen in **example 26** and **69**, there still seems to be confusion in distinguishing the issue types of *Lexical Selection* and *Overly Literal*, as shown in **examples 70-75**.

Annotation of mistranslation	
Unbabel Error Typology	East Asian Languages Annotation Module
EN (source) (24a) No worries. JA (target) (24b) <u>ご心配には及びません。</u> [A: Overly Literal] [B: Lexical Selection]	EN (source) (69a) No worries. JA (target) (69b) <u>ご心配には及びません。</u> [A: Overly Literal] [B: Overly Literal]

Table 50. Example of EN-JA annotation of *Mistranslation* with the East Asian Languages Annotation Module for the Unbabel Quality Framework

Annotation of mistranslation	
Annotator A	Annotator B
<p>EN (source) (70a) We have successfully cancel the automatic payment under billing ID ALPHANUMERICID-0</p> <p>JA (target) (70b) 請求ID ALPHANUMERICID-0の下にある自動支払いを正常にキャンセルいたしました。 [Overly Literal, major]</p>	<p>EN (source) (71a) We have successfully cancel the automatic payment under billing ID ALPHANUMERICID-0</p> <p>JA (target) (71b) 請求ID ALPHANUMERICID-0の下にある自動支払いを正常にキャンセルいたしました。 [Lexical Selection, major]</p>
<p>EN (source) (72a) Nice to meet you.</p> <p>ZH-TW (target) (72b) 很高興見到您。 [Lexical Selection, major]</p>	<p>EN (source) (73a) Nice to meet you.</p> <p>ZH-TW (target) (73b) 很高興見到您。 [Overly Literal, major]</p>
<p>EN (source) (74a) Please give me one moment here.</p> <p>ZH-CN (target) (74b) 请给我片刻。 [Overly Literal, minor]</p>	<p>EN (source) (75a) Please give me one moment here.</p> <p>ZH-CN (target) (75b) 请给我片刻。 [Lexical Selection, minor]</p>

Table 51. Examples of EN-JA, EN-ZH_TW and EN-ZH_CN annotation of *Mistranslation* with the East Asian Languages Annotation Module for the Unbabel Quality Framework

Similarly, as seen in **examples 76** and **77**, there is still confusion regarding *Mistranslation* and other issue types outside of it, such as *MT Hallucination*.

Issue type disagreement	
Annotator A	Annotator B
EN (source) (76a) Korean is not working earlier. KO (target) (76b) 중국어가 이전에는 작동하지 않습니다. [Lexical Selection, critical]	EN (source) (77a) Korean is not working earlier. KO (target) (77b) 중국어가 이전에는 작동하지 않습니다. [MT Hallucination, critical]

Table 52. Example of EN-KO issue type disagreement with the East Asian Languages Annotation Module for the Unbabel Quality Framework

Since the *Tricky Cases* section of the guidelines corresponding to this annotation module was lengthy and detailed in relation to annotation of verbs, as shown earlier in **Table 17**, it was interesting to analyze how the annotators performed in relation to this. The first issue pointed out in the guidelines, which was observed through the analysis of several datasets, was that annotators for Japanese and Korean had been annotating *Tense/Mood/Aspect* one one single character instead of a whole unit. However, as there were no such errors in all three rounds of annotations for the experiments it was concluded that this problem has been mostly resolved through the feedback provided to the annotators, as mentioned in **Section 4.2.**

The second issue that was pointed out was the span selection of verbs, particularly for Chinese. As mentioned before, tense, mood and aspect in Chinese is expressed through the usage of elements such as particles or adverbs. As such, it was thought that the best way of annotating these errors in Chinese would be to allow multi-selection so that the annotators could select all parts of the sentence that were related to the verb being in the wrong form. While this was successfully applied while annotating verbs in Chinese, as seen in **example 78** and **79**, *Annotator B* for Korean also unexpectedly used the multi-selection function, which caused disagreement as *Annotator A* annotated the same errors without this function, as seen in **example 80** and **81**.

Annotation of Tense/Mood/Aspect	
Unbabel Error Typology	East Asian Languages Annotation Module
EN (source) (78a) Okay, may I know when and where did you purchase your node?	EN (source) (79a) Okay, may I know when and where did you purchase your node?
ZH-CN (target) (78b) 好的, 请问您何时何地购买节点? [Other POS Omitted, major] [Tense/Mood/Aspect, major]	ZH-CN (target) (79b) 好的, 请问您何时何地购买[Ø]节点? (购买 + [Ø]) [Tense/Mood/Aspect, major]

Table 53. Example of annotation of *Tense/Mood/Aspect* in Simplified Chinese with the East Asian Languages Annotation Module for the Unbabel Quality Framework

Annotation of Tense/Mood/Aspect	
Annotator A	Annotator B
EN (source) (80a) What do the LEDs on the [PRODUCT] indicate?	EN (source) (81a) What do the LEDs on the [PRODUCT] indicate?
KO (target) (80b) [PRODUCT]의 LED는 무엇을 나타냅니다. [Tense/Mood/Aspect, major]	KO (target) (81b) [PRODUCT]의 LED는 무엇을 나타냅니다. (무엇을 + 나타냅니다) [Tense/Mood/Aspect, major]

Table 54. Annotation of *Tense/Mood/Aspect* in Korean with the East Asian Languages Annotation Module for the Unbabel Quality Framework

As illustrated in **examples 82-89**, in the case of Span and Severity, which were addressed in the *Tricky Cases* and *Decision Trees* sections of the guidelines respectively, although the guidelines were effective in some cases, there are still instances of disagreement between the annotators.

Agreement on span and severity	
Unbabel Error Typology	East Asian Languages Annotation Module
<p>EN (source) (82a) Their email services are still active.</p> <p>JA (target) (82b) メールサービスはまだ有効です。 [Other POS Omitted, major] [A:メール] [B:メールサービス]</p>	<p>EN (source) (83a) Their email services are still active.</p> <p>JA (target) (83b) メールサービスはまだ有効です。 [Other POS Omitted, major] [A:メール] [B:メール]</p>
<p>EN (source) (84a) We can provide you the default firmware, but for you to remove the current one, you have to contact the firmware provider.</p> <p>ZH-CN (target) (84b) 我们可以为您提供默认固件，但是为了删除当前的 one，您必须与固件提供商联系。 [A: Untranslated, critical] [B: Untranslated, major]</p>	<p>EN (source) (85a) We can provide you the default firmware, but for you to remove the current one, you have to contact the firmware provider.</p> <p>ZH-CN (target) (85b) 我们可以为您提供默认固件，但是为了删除当前的 one，您必须与固件提供商联系。 [A: Untranslated, major] [B: Untranslated, major]</p>

Table 55. Example of EN-JA and EN-ZH_CN agreement on Span and Severity with the East Asian Languages Annotation Module for the Unbabel Quality Framework

Disagreement on span and severity	
Annotator A	Annotator B
<p>EN (source) (86a) Upon checking, the card went missing because it was disabled.</p> <p>JA (target) (86b) 確認いたしましたところ、カードは無効になっておりますので、足りなくなっております。 [Lexical Selection, major]</p>	<p>EN (source) (87a) Upon checking, the card went missing because it was disabled.</p> <p>JA (target) (87b) 確認いたしましたところ、カードは無効になっておりますので、足りなくなっております。 [Lexical Selection, major]</p>
<p>EN (source) (88a) If you still unable to make payment, please contact your card issuer.</p> <p>JA (target) (88b) それでもお支払いができない場合は、カード発行者にお問い合わせください。 [Lexical Selection, major]</p>	<p>EN (source) (89a) If you still unable to make payment, please contact your card issuer.</p> <p>JA (target) (89b) それでもお支払いができない場合は、カード発行者にお問い合わせください。 [Lexical Selection, minor]</p>

Table 56. Example of EN-JA disagreement on Span and Severity with the East Asian Languages Annotation Module for the Unbabel Quality Framework

It is arguable that severity selection, and even span selection in some cases, may be susceptible to a higher degree of subjectivity than the selection of issue types and, thus, it cannot be said that one annotator is wrong and the other is right. However, some disagreement in the annotations while using this annotation module could have been avoided if the annotators had adhered to the guidelines completely.

Issue type disagreement	
Annotator A	Annotator B
Source (90a) 支払いは競合の影響を受けませんのでご安心ください。	Source (91a) 支払いは競合の影響を受けませんのでご安心ください。
JA (target) (90b) 支払い橋合同教室を慈善団体楽天会 [Do not Translate, critical]	JA (target) (91b) 支払い橋合同教室を慈善団体楽天会 [MT Hallucination, critical]

Table 57. Example of EN-JA issue type disagreement with the East Asian Languages Annotation Module for the Unbabel Quality Framework

In **example 90** and **91**, the source text was already in the intended target language but it was still translated into a different target text. As per the definition provided in the guidelines, the correct issue type to use for this error was *Do not Translate*. However, *Annotator B* used the *MT Hallucination* issue type.

<p>Do not Translate</p> <p>Text was translated and it should have been left untranslated. Please check if it's a brand or any other foreign word, and follow your language rules regarding those cases.</p> <p>In case the source text is already written in the target language and it has been altered in the target, either to a completely different meaning or to paraphrasing, this error type should be used.</p> <p>Ex:</p> <p>(Source: ZH-CN) 感谢您提供详细信息。</p> <p>(Target: ZH-CN) [请您告诉我们今天的表现如何]DO NOT TRANSLATE。</p> <p>(Target: ZH-CN) [感谢您提供详细信息]DO NOT TRANSLATE。</p>

Table 58. Definition of *Do not Translate* errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework

In addition, the annotators who previously ignored wrong punctuation usage, particularly in the case of Simplified Chinese and full-width punctuation, continued to annotate this wrongly even though the correct way to do so was clearly stated in the guidelines.

Punctuation

Punctuation is used incorrectly or is missing, or one of a pair of quotes, brackets or punctuation is missing from the target text. **Please note** that punctuation that is supposed to be full-width, as is the case for commas in Chinese and Japanese, should also be annotated using this tag if they don't appear as such in the target.

Ex:

(EN) Thank you for providing a screenshot.

(KO) 스크린샷[.]PUNCTUATION를 제공해 주셔서 감사합니다.

(KO) 스크린샷[Ø]PUNCTUATION를 제공해 주셔서 감사합니다.

Table 59. Definition of *Punctuation* errors in the East Asian Languages Annotation Module for the Unbabel Quality Framework

In the Simplified Chinese annotations, where *Annotator B* had previously ignored all wrong punctuation errors that had to do with punctuation marks being in half-width, while using the East Asian Languages Annotation Module for the Unbabel Quality Framework they annotated these errors. However, they did it wrongly as they used the *Whitespace* issue type on the words immediately following the punctuation in the wrong format.

Annotation of punctuation in Simplified Chinese	
	<p>Whitespace</p> <ul style="list-style-type: none"> ● "遗憾"
	<p>Whitespace</p> <ul style="list-style-type: none"> ● "您"
	<p>Whitespace</p> <ul style="list-style-type: none"> ● "因为"

Table 60. Example of annotation of *Punctuation* in Simplified Chinese with the East Asian Languages Annotation Module for the Unbabel Quality Framework

Finally, it is important to point out that this annotation module also produced positive results in relation to some of the changes that were introduced by the newest version of the Unbabel Error Typology (v3). The most notorious cases are those of *Register* and *Terminology* due to the fact that both were simplified. In the case of *Register* this became a single issue type with no further categorization, while *Terminology* went from being subdivided into three issue types to two. As seen in **examples 92 to 95**, this was helpful in raising agreement between annotators.

Agreement on register and terminology	
Unbabel Error Typology	East Asian Languages Annotation Module
EN (source) (92a) Hello there how are you doing? KO (target) (92b) 안녕 잘 지냈어 요? [A: Lexical Register, minor] [B: Grammatical Register, minor]	EN (source) (93a) Hello there how are you doing? KO (target) (93b) 안녕 잘 지냈어 요? [A: Register, minor] [B: Register, minor]
EN (source) (94a) We love to help you remove it, but we don't have the resources to support this [PRODUCT] since the router is an open source. ZH-CN (target) (94b) 我们很乐意帮助您移除它, 但由于路由器是开放的 zh-CN , 因此我们没有支持此 [PRODUCT] 的资源。 [A: Mistranslated Term, major] [B: Term Wrongly Applied, major]	EN (source) (95a) We love to help you remove it, but we don't have the resources to support this [PRODUCT] since the router is an open source. ZH-CN (target) (95b) 我们很乐意帮助您移除它, 但由于路由器是开放的 zh-CN , 因此我们没有支持此 [PRODUCT] 的资源。 [A: Wrong Term, major] [B: Wrong Term, major]

Table 61. Examples of EN-KO and EN-ZH_CN agreement on *Register* and *Terminology* with the East Asian Languages Annotation Module for the Unbabel Quality Framework

5.3.3. Annotator’s Feedback

As mentioned before on **Section 4.3.2.2.**, aside from a comment from one of the Japanese annotators asking for clarification about the *Tricky Cases* section about annotation with particles, the annotators did not provide any feedback or any other questions regarding the annotation module. This could be due partly to the fact that the East Asian Languages Annotation Module for the Unbabel Quality Framework addresses a large part of the problems previously pointed out by the annotators in relation to the Unbabel Error Typology. In addition to the issue types that

were removed from the typology, which were explained in **Section 4.3.2.**, some of the issue types proposed by the annotators were included in this annotation module, as is the case of *Omitted Particle*, *Omitted Classifier* and *Wrong Classifier*. At the same time, however, it is important to note that not all the categories which are demonstrated in **Table 31**, suggested by the annotators, were integrated into the East Asian Languages Annotation Module. Aside from the proposed issue type of *Inappropriate*, which in the definition provided by the annotator was very similar to the *Overly Literal* issue type, three other *Omission* issue types were not included. This was due to the fact that it was believed that these issue types would make the annotation module overly fine-grained and, based on previous analysis of annotation data, there were not enough translation errors related to these categories that would justify their inclusion in the annotation module, which is further proved by the fact that the annotators provided no feedback regarding missing categories within the East Asian Languages Annotation Module.

In conclusion, based on the results that were presented in this chapter, it is possible to affirm that the East Asian Languages Annotation Module for the Unbabel Quality Framework which is proposed in this thesis has both strengths and limitations which can serve as a basis for its future improvement. This annotation module was not able to entirely solve the divergence between annotators in relation to errors existing under *Mistranslation*. As errors existing under this category are among the most frequent, the fact that there is visible disagreement between annotators over which issue type should be attributed for each case of mistranslation has a heavy negative impact on IAA scores. In addition, lack of training of the annotators and clarity in the guidelines also proved to have a negative reflection on IAA scores and are also areas that need to be addressed.

On the other hand, however, the annotation module was successful in implementing the issue types related to particles and classifiers which had been previously requested by the annotators in past feedback. While the annotators also made use of the equivalent issue types when annotating with the typology proposed by Ye and Toral (2020), the final result was more positive with the East Asian Languages Annotation Module for the Unbabel Quality Framework due to the organization of the typology itself, as shown in the examples provided. In relation to the newly introduced *Transliteration* issue type, since its implementation was not as successful it

remains to be investigated whether it would be of more benefit to remove this issue type in future versions of the annotation module or to improve the way it is defined in the guidelines.

6. Conclusions and Future Work

The objective of this thesis was to investigate how using a translation error typology that is inadequate for a certain content type or, in this case, for a specific group of languages, for annotation can influence the reliability of the annotations as well as the agreement between annotators whose work suffers from levels of ambiguity and subjectivity higher than desired. In light of this, during the internship at Unbabel, an annotation module for East Asian languages was developed with the aim of obtaining more consistency in annotations and a more accurate reflection of the true quality of the translations for these languages. As such, in addition to explaining the consequences of using error typologies that are not a good fit for the content that is being annotated with concrete and real examples, this dissertation also aimed to explore the benefits of annotating with a typology dedicated to an exclusive set of languages.

This project was proposed due to the fact that there was an internal need to adapt Unbabel's quality processes to the languages under discussion and that there was concrete feedback from annotators regarding the points that should be improved. As discussed in **Section 4.3.1.**, in this thesis, there was a large amount of wrong annotations in previous data, many of which could be attributed to the lack of appropriate issue types for some errors and confusion regarding the annotation guidelines, which could also be related to the existence of unnecessary issue types in relation to the languages under analysis.

The process of building the East Asian Languages Annotation Module for the Unbabel Quality Framework implied both adding and removing issue types regarding the Unbabel Error Typology, in order to create a separate taxonomy that was relevant for annotation of the languages at hand. While removing issue types such as *Agreement* served the purpose of removing unnecessary noise by eliminating issue types that are not used when annotating these languages, adding specific issue types for components relevant to these languages aimed to improve agreement between annotators by providing issue types to cover specific language components whose annotation was previously unclear, as had already been pointed out by the annotators.

In order to obtain a comprehensive comparison, in addition to the annotation module that was built in the scope of this thesis and the Unbabel Error Typology, a round of annotations was also conducted using the MQM-compliant error taxonomy for the translation direction of English

to Chinese proposed by Ye and Toral (2020). This experiment was carried out over the span of three months, allowing the annotators one month to annotate each batch. However, it was possible to identify a progressive level of exhaustion in the annotators after the first batch.

The results demonstrated that all error typologies tested had strengths and limitations. In the case of the Unbabel Error Typology, while it benefitted from the fact that it was a typology familiar to the annotators, overall its performance was not overly superior than that of the other two typologies in terms of IAA, due to it being the most extensive and fine-grained typology out of the three, with a total of 47 selectable issue types and some, as previously stated were not even relevant to the languages analyzed. In addition, the lack of the issue types that had been previously suggested by the annotators also had a negative influence on the IAA scores because it forces the annotators to make relatively subjective decisions which are less likely to be in agreement with those of other annotators.

On the other hand, the typology proposed by Ye and Toral (2020) had the advantage of being conceived specifically for annotation of target text in Chinese. This means that the typology contains some of the issue types that do not exist in the Unbabel Error Typology and which were considered important by the annotators, namely issue types related to particles and classifiers. In addition, this typology is much simpler than the other two under comparison, with a total of 15 selectable issue types. Although these were all characteristics that were expected to have a positive impact on IAA, this was not verified fully. One of the biggest limitations of this typology was the organization of the issue types under *Function Words*, which was interpreted differently by the annotators, despite the annotation guidelines they were provided with. Furthermore, the lack of categories under *Mistranslation* also generated much disagreement between annotators. In relation to this typology, the annotators also considered that the lack of issue types for *Register* and *Whitespace* severely affected the annotation process. As indicated in Ye and Toral (2020), in their study an average IAA score of 0.44 Cohen's kappa was obtained when annotating with this typology. During the annotation experiments performed for this thesis, the typology proposed by Ye and Toral (2020) scored similarly for the LP of English to Korean and higher for English to Japanese. Interestingly enough, the lower IAA scores correspond to both varieties of Chinese. As such, it is important to note that although this typology was conceived for annotation of Chinese, it performed comparatively better for Korean, as this was the LP for which the typology proposed by Ye and Toral (2020) visibly had IAA scores higher

than the other two typologies, which indicates that this typology can be transversal to LPs outside of English-Chinese.

Although the East Asian Languages Annotation Module for the Unbabel Quality Framework which was proposed on this thesis did not demonstrate results that were superior to those obtained with the other typologies, a detailed analysis of the annotations revealed that its implementation was mostly successful, which is visible from the fact that it is the typology with the highest percentage of jobs scored above the 0.4 threshold in terms of IAA. The East Asian Languages Annotation Module is structurally similar to the Unbabel Error Typology, which had the benefit of not being foreign to the annotators, but it is more simplified. This simplification reduced some of the previous agreement problems in relation to issue types such as *Register* and *Terminology* but, similarly to the Unbabel Error Typology, it continued to demonstrate weaknesses in relation to *Mistranslation*, including the newly proposed issue type of *Transliteration*. On the other hand, however, the implementation of the issue types that had been requested by annotators in relation to particles and classifiers was very successful, which demonstrates that this proposal still needs to be perfected in future iterations of the annotation module, but that it already has a good foundation.

It is proposed that this annotation module is integrated in the Unbabel Quality Framework as a language-specific annotation module that would be complementary to a set of issue types common to all languages. This would consist of a new approach to annotation, where there would exist a core typology consisting of issue types that are applicable to all languages and additional annotation modules that would be dedicated to specific languages or language groups, thus containing issue types that are exclusive to certain languages and not relevant for others. This is the fundamental reason for our designation of the East Asian Language typology subset as a Module.

As mentioned before, it will be necessary to keep further improving the application of the East Asian Languages Annotation Module for the Unbabel Quality Framework, both through improving the annotation module itself and training the annotators. Furthermore, it will be important to investigate if this module is equally transversal to all the languages which have been analyzed or if it would be necessary to separate them. Similarly, since the experiments of this thesis were conducted using only chat data, the compatibility of this annotation module with other content types remains to be analyzed. Finally, it is expected that in the future this

annotation module can be integrated into automatic workflows for quality evaluation, both in terms of semi-automatic human evaluation and quality metrics.

7. Bibliography

- Akamatsu, T. (2011). Honorific Particles in Japanese and Personal Monemes. *Presses Universitaires de France*, 47(1), 37–49.
- Amidei, J., Piwek, P., & Willis, A. (2019). Agreement is overrated: A plea for correlation to assess human evaluation reliability. *Proceedings of the 12th International Conference on Natural Language Generation*, 344–354. <https://doi.org/10.18653/v1/W19-8642>
- Artstein, R. (2017). *Handbook of Linguistic Annotation* (N. Ide & J. Pustejovsky, Eds.; 1st ed. 2017). Springer Netherlands: Imprint: Springer. <https://doi.org/10.1007/978-94-024-0881-2>
- Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. <https://aclanthology.org/W05-0909.pdf>
- Bunt, J. (2003). *The Oxford Japanese grammar and verbs* (1. publ). Oxford Univ. Press.
- Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to Human and Machine Translation Quality Assessment. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation Quality Assessment* (Vol. 1, pp. 9–38). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_2
- Castilho, S., Gaspari, F., Moorkens, J., Popović, M., & Toral, A. (2019). Editors' foreword to the special issue on human factors in neural machine translation. *Machine Translation*, 33(1–2), 1–7. <https://doi.org/10.1007/s10590-019-09231-y>

- Chida, S. (2015). Analysis of Japanese Particle Errors Made by Marathi JSL Learners. *Vice Chancellor, Deccan College Post-Graduate and Research Institute (Deemed University)*, 75, 215–224.
- Cho, W. I., Moon, S., & Song, Y. (2020). Open Korean Corpora: A Practical Report. *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, 85–93. <https://doi.org/10.18653/v1/2020.nlposs-1.12>
- Fang, J., & Connelly, M. (Eds.). (2008). *Cheng & Tsui Chinese measure word dictionary: A Chinese-English English-Chinese usage guide* = . Cheng & Tsui Company.
- Gino, D. (2019, April 23). *The State of Neural Machine Translation for Asian Languages*. Slator. <https://slator.com/the-state-of-neural-machine-translation-for-asian-languages/>
- Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2013). Continuous Measurement Scales in Human Evaluation of Machine Translation. *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, 33–41. <https://aclanthology.org/W13-2305.pdf>
- Hutchins, J. (2003). *Machine Translation: A Concise History*. 1–21.
- Ishisaka, T., Utiyama, M., Sumita, E., & Yamamoto, K. (2009). *Development of a Japanese-English Software Manual Parallel Corpus*. Proceedings of Machine Translation Summit XII: Posters. <https://aclanthology.org/2009.mtsummit-posters.10.pdf>
- Kepler, F., Trénous, J., Treviso, M., Vera, M., & Martins, A. F. T. (2019). *OpenKiwi: An Open Source Framework for Quality Estimation* (arXiv:1902.08646). arXiv. <http://arxiv.org/abs/1902.08646>

- Koehn, P. (2020). *Neural Machine Translation* (1st ed.). Cambridge University Press.
<https://doi.org/10.1017/9781108608480>
- Koo, J. H. (1997). Measure Words: Native Korean and Sino-Korean. *Penn State University Press*, 2, 193–204.
- Lee, H.-G., Lee, J., Kim, J.-S., & Lee, C.-K. (2015). NAVER Machine Translation System for WAT 2015. *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, 69–73. <https://aclanthology.org/W15-5008.pdf>
- Lee, S.-H., Jang, S. B., & Seo, S. (2009). Annotation of Korean Learner Corpora for Particle Error Detection. *Equinox Publishing Ltd.*, 26(3), 529–544.
- Li, C. N., & Thompson, S. A. (2009). *Mandarin Chinese: A functional reference grammar* (1. paperback print, repr). Univ. of California Press.
- Loh, S., & Kong, L. (1977). Computer Translation of Chinese Scientific Journals. *München: Verlag Dokumentation*, 1.
<https://aclanthology.org/www.mt-archive.info/CEC-1977-Loh.pdf>
- Loh, S., & Kong, L. (1979). An interactive on-line machine translation system (Chinese into English). *North-Holland Publishing Company*. <https://aclanthology.org/1978.tc-1.7.pdf>
- Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica: Tecnologies de La Traducció*, 12, 455.
<https://doi.org/10.5565/rev/tradumatica.77>
- Lommel, A. (n.d.). Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In *Translation Quality Assessment* (Vol. 1, pp. 109–127). Springer International Publishing. <https://link.springer.com/book/10.1007/978-3-319-91241-7>

- Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (1st ed., pp. 135–158). Cambridge University Press.
<https://doi.org/10.1017/CBO9781139649414.007>
- Makino, S., & Tsutsui, M. (1989). *A dictionary of basic Japanese grammar =: Nihongo kihon bunpō jiten* (1st pbk. ed). Japan Times.
- Martin, S. E. (1992). *A reference grammar of Korean: A complete guide to the grammar and history of the Korean language* (1st. ed). Charles E. Tuttle.
- Martins, A. F. T., Junczys-Dowmunt, M., Kepler, F. N., Astudillo, R., Hokamp, C., & Grundkiewicz, R. (2017). Pushing the Limits of Translation Quality Estimation. *Transactions of the Association for Computational Linguistics*, 5, 205–218.
https://doi.org/10.1162/tacl_a_00056
- MQM Core Typology*. (n.d.). Retrieved 18 December 2021, from <https://themqm.info/typology/>
- MQM Definition*. (n.d.). QT21. Retrieved 15 November 2021, from <https://www.qt21.eu/mqm-definition/definition-2015-12-30.html>
- MQM (Multidimensional Quality Metrics)*. (n.d.). <https://themqm.org/>
- Nagao, M. (1989). Japanese view of the future of machine translation. *Proceedings of Machine Translation Summit II*. <https://aclanthology.org/1989.mtsummit-1.18.pdf>
- Official Journal of The European Union*. (2016, May 4). <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN>

- Pak, M. D. (2008). Types of Clauses and Sentence end Particles in Korean. *Korean Linguistics*, 14, 113–156. <https://doi.org/10.1075/kl.14.06mdp>
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. <https://doi.org/10.3115/1073083.1073135>
- Park, S.-Y., & Oh, G.-R. (1999). *Machine translation in Korea*. 100–106. <https://aclanthology.org/www.mt-archive.info/MTS-1999-Park-1.pdf>
- Raunak, V., Menezes, A., & Junczys-Dowmunt, M. (2021). The Curious Case of Hallucinations in Neural Machine Translation. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1172–1183. <https://doi.org/10.18653/v1/2021.naacl-main.92>
- Rawling, P., & Wilson, P. (Eds.). (2019). *The Routledge handbook of translation and philosophy*. Routledge Taylor & Francis Group.
- Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. <https://doi.org/10.18653/v1/2020.emnlp-main.213>
- Specia, L., Scarton, C., & Paetzold, G. H. (2018). *Quality Estimation for Machine Translation*. Springer International Publishing. <https://doi.org/10.1007/978-3-031-02168-8>
- Sun, M., Jiang, B., Xiong, H., He, Z., Wu, H., & Wang, H. (2019). Baidu Neural Machine Translation Systems for WMT19. *Proceedings of the Fourth Conference on Machine*

- Translation (Volume 2: Shared Task Papers, Day 1)*, 374–381.
<https://doi.org/10.18653/v1/W19-5341>
- Susanto, R. H., Wang, D., Yadav, S., Jain, M., & Htun, O. (2021). Rakuten’s Participation in WAT 2021: Examining the Effectiveness of Pre-trained Models for Multilingual and Multimodal Machine Translation. *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, 96–105. <https://doi.org/10.18653/v1/2021.wat-1.9>
- Tang, S.-W. (2015). A generalized syntactic schema for utterance particles in Chinese. *Lingua Sinica*, 1(1), 3. <https://doi.org/10.1186/s40655-015-0005-5>
- Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, O., Lu, Y., Li, S., Wang, Y., & Wang, L. (2014). *UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation*. 1837–1842.
http://www.lrec-conf.org/proceedings/lrec2014/pdf/774_Paper.pdf
- Toral, A., & Sánchez-Cartagena, V. M. (2017). *A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions*.
<https://doi.org/10.48550/ARXIV.1701.02901>
- Transliteration. (2022). In *Cambridge English Dictionary*.
<https://dictionary.cambridge.org/dictionary/english/transliteration>
- Wang, Y., Cheng, S., Jiang, L., Yang, J., Chen, W., Li, M., Shi, L., Wang, Y., & Yang, H. (2017). Sogou Neural Machine Translation Systems for WMT17. *Proceedings of the Second Conference on Machine Translation*, 410–415.
<https://doi.org/10.18653/v1/W17-4742>
- What is CRM?* (n.d.). Salesforce. Retrieved 10 September 2022, from
<https://www.salesforce.com/eu/learning-centre/crm/what-is-crm/>

WMT Conference on Machine Translation. (n.d.). Machine Translate. Retrieved 8 December 2021, from <https://machinetranslate.org/wmt>

Wu, S., Wang, X., Wang, L., Liu, F., Jun, J., Tu, Z., Shi, S., & Li, M. (2020). *Tencent Neural Machine Translation Systems for the WMT20 News Translation Task*. 313–319. <https://aclanthology.org/2020.wmt-1.34.pdf>

Ye, Y., & Toral, A. (2020). *Fine-grained Human Evaluation of Transformer and Recurrent Approaches to Neural Machine Translation for English-to-Chinese* (arXiv:2006.08297). arXiv. <http://arxiv.org/abs/2006.08297>

8. Annexes

A. Issue Types per Typology

	Unbabel Error Typology	Typology proposed by Ye and Toral	East Asian Languages Annotation Module
Addition	✓	✓	✓
Omission		✓	
Omitted Preposition	✓	✗	✓
Omitted Conjunction	✓	✗	✓
Omitted Determiner	✓	✗	✗
Omitted Pronoun	✓	✗	✓
Omitted Auxiliary Verb	✓	✗	✓
Other POS Omitted	✓	✗	✓
Omitted Particle	✗	✗	✓
Omitted Classifier	✗	✗	✓
Mistranslation			
Ambiguous Translation	✓	✗	✗
Named Entity	✓	✓	✓
False Friend	✓	✗	✗
Overly Literal	✓	✓	✓
Lexical Selection	✓	✗	✓

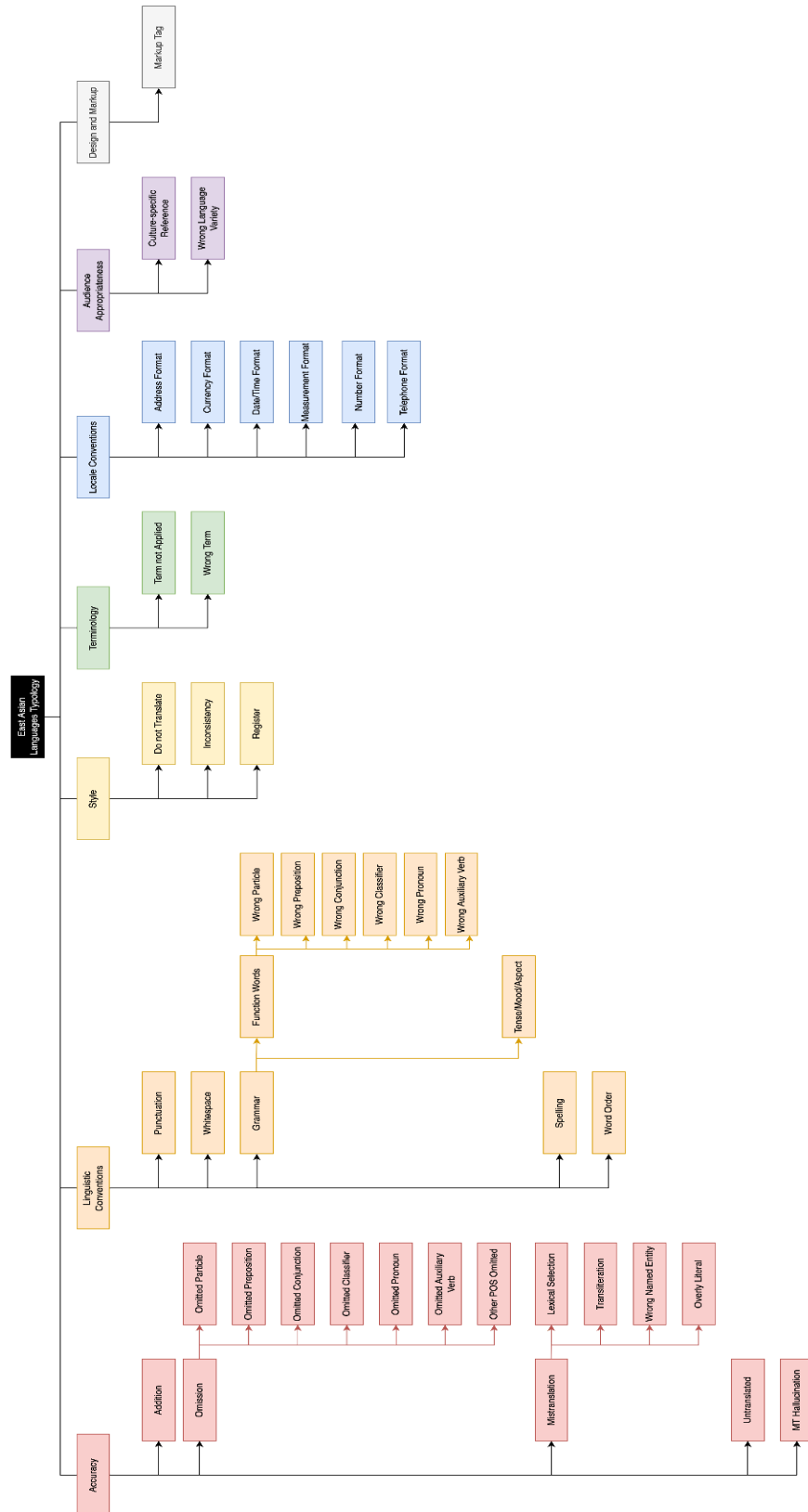
Shouldn't Have Been Translated	✓	✗	✓
Spelling		✗	✓
<i>Source/Target Disagreement</i>	✓	✗	✗
<i>Wrong Paronym</i>	✓	✗	✗
Wrong Date/Time	✓	✗	✗
Wrong Number	✓	✗	✗
Wrong Unit Conversion	✓	✗	✗
Transliteration	✗	✗	✓
Untranslated	✓	✓	✓
MT Hallucination	✓	✗	✓
Character Encoding	✓	✗	✗
Duplication	✓	✗	✗
Grammar			
Function Words			
<i>Wrong Preposition</i>	✓	✗	✓
<i>Wrong Conjunction</i>	✓	✗	✓
<i>Wrong Determiner</i>	✓	✗	✗
<i>Wrong Pronoun</i>	✓	✗	✓
<i>Wrong Auxiliary Verb</i>	✓	✗	✓
<i>Wrong Particle</i>	✗	✗	✓
<i>Wrong Classifier</i>	✗	✗	✓

<i>Extraneous</i>	×		×
- <i>Preposition</i>	×	✓	×
- <i>Adverb</i>	×	✓	×
- <i>Particle</i>	×	✓	×
<i>Incorrect</i>	×	✓	×
<i>Missing</i>	×	✓	×
Classifier	×	✓	×
Word Form		×	×
<i>Agreement</i>	✓	×	×
<i>Tense/Mood/Aspect</i>	✓	×	✓
<i>Part of Speech</i>	✓	×	×
Word Order	✓	✓	✓
Inconsistency	✓	×	✓
Typography			×
Capitalization	✓	×	×
Diacritics	✓	×	×
Hyphenation	✓	×	×
Orthography	✓	×	×
Punctuation	✓	✓	✓
Whitespace	✓	×	✓
Unpaired-marks	×	✓	×

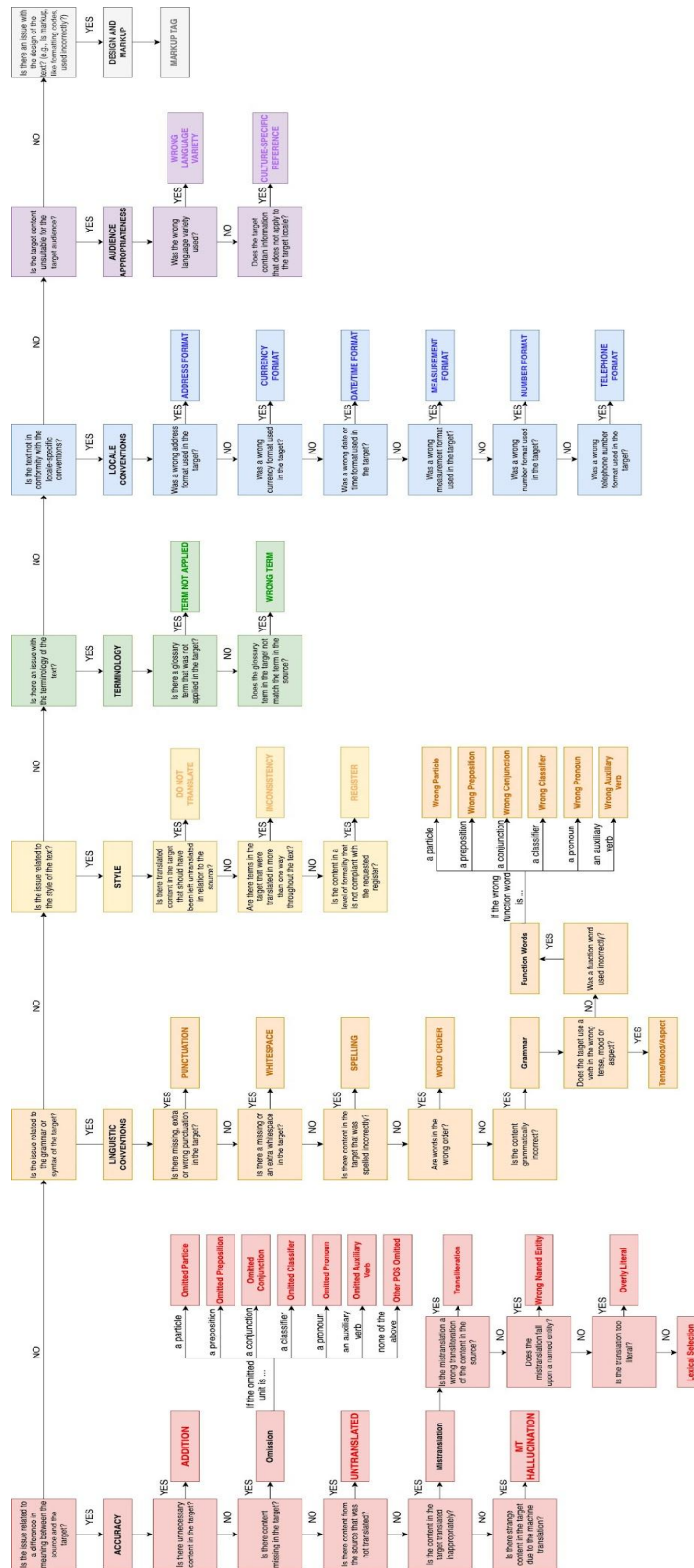
Unintelligible	✓	✓	✗
Register		✗	✓
Grammatical Register	✓	✗	✗
Lexical Register	✓	✗	✗
Wrong Language Variety	✓	✗	✓
Meaning Relations		✗	✗
Hyperonym/Hyponym	✓	✗	✗
Synonym	✓	✗	✗
Company Terminology		✗	
Term not Applied	✓	✗	✓
Mistranslated Term	✓	✗	✗
Term Wrongly Applied	✓	✗	✗
Wrong Term	✗	✗	✓
Address Format	✗	✗	✓
Currency Format	✗	✗	✓
Date/Time Format	✗	✗	✓
Measurement Format	✗	✗	✓
Number Format	✗	✗	✓
Telephone Format	✗	✗	✓
Culture-specific Reference	✗	✗	✓

Markup Tag	✘	✘	✔
TOTAL OF SELECTABLE ISSUE TYPES	47	15	39

B. East Asian Languages Annotation Module for the Unbabel Quality Framework



C. Decision Tree for issue type selection in the East Asian Languages Annotation Module for the Unbabel Quality Framework



D. Decision Tree for severity selection in East Asian Languages Annotation Module for the Unbabel Quality Framework

