

UNIVERSIDADE DE LISBOA
FACULDADE DE LETRAS



**Analysis on the impact of the source text quality:
Building a data-driven typology**

Madalena Sofia Nunes Gonçalves

Orientado pela Professora Doutora Helena Gorete Silva Moniz e pela Mestre Marianna Buchicchio, especialmente elaborada para a obtenção do grau de mestre em Tradução, na modalidade de relatório de estágio

2021

Dedicatória

*Para o meu avô Nunes,
que me ensinou que o trabalho
árduo compensa sempre.*

Acknowledgments

Às minhas orientadoras, Prof^a. Doutora Helena Moniz e Mestre Marianna Buchicchio por todo o apoio durante este estágio.

À Marianna, nem sei como te agradecer por tudo o que fizeste por mim. Para além de uma excelente coordenadora que sempre se mostrou disponível e que me motivou a ter mais iniciativa no mundo do trabalho, também te tornaste uma amiga.

À Professora Helena, que me introduziu ao mundo da Tradução Automática e que me incentivou a não receá-lo. A constante proteção e honestidade foram essenciais para esta fase da minha vida.

À Unbabel, especialmente as equipas em que me inseri *Linguistic Services* e *MT Team*. Fui recebida com a melhor hospitalidade possível.

Aos meus amigos, que embora longe estiveram sempre presentes e me ajudaram a relaxar. Vocês sabem quem são.

À minha irmã, que aguenta o meu lado chato, exaltado e stressado, mas que raramente se queixa.

Aos meus pais, que são excelentes modelos e que tornaram a minha educação académica possível. Sem o vosso apoio nunca estaria onde estou agora, feliz e otimista.

Index

Abstract	8
Resumo	9
1. Introduction	13
2. Host characterization	15
2.1 Content types	15
2.2 Unbabel translation pipeline	16
2.2.1 Tickets translation pipeline	16
2.2.2 Chat translation pipeline	17
2.2.3 FAQs translation pipeline	18
2.3 Unbabel Quality processes	18
2.3.1 Human Evaluation	19
2.4 Evaluation process: evaluating the post-editors' community	22
3. State of the art	23
3.1 Machine Translation	23
3.2 Translation Quality Evaluation	26
3.2.1 Automatic quality metrics	27
3.2.2 Manual quality metrics	29
3.3 Error Typologies: strengths and limitations	32
4. Methodology	37
4.1 Project's objective	38
4.2 Building a data driven typology	38
4.2.1 Agent annotation EN	40
4.2.2 User annotation PT-BR	43
4.2.3 Source Typology: proposed issue types	47

4.2.3.1 Annotation guidelines	69
4.2.3.1.1 Span types	70
4.2.3.1.2 Tricky cases	75
4.2.3.1.3 Severities	82
4.2.3.1.4 Decision trees	85
5. Results and discussion	88
5.1 PT-BR_EN inbounds	88
5.2 Agent Annotation	95
5.3 Multilingual internal pilot	99
5.3.1 Neutral structures analysis	101
5.3.2 Critical errors analysis	114
5.3.3 Typology Misusage	116
5.3.4 Inter Annotator Agreement	120
5.3.4.1 Study of IAA of Internal pilot	122
6. Language eGuides	127
7. Conclusions and Future work	129
Bibliography	132
Annexes	143

Índice

Abstract	8
Resumo	9
1. Introdução	13
2. Caracterização da entidade de acolhimento	15
2.1 Tipos de Conteúdo	15
2.2 Processos de tradução da Unbabel	16
2.2.1 Processo de tradução de <i>Tickets</i>	16
2.2.2 Processo de tradução de <i>Chat</i>	17
2.2.3 Processo de tradução de <i>FAQs</i>	18
2.3 Processos de Qualidade da Unbabel	18
2.3.1 Avaliação Humana	19
2.4 Processo de avaliação: avaliar a comunidade de pós-editores	22
3. Estado da Arte	23
3.1 Tradução Automática	23
3.2 Avaliação da Qualidade de Tradução	26
3.2.1 Métricas de qualidade automáticas	27
3.2.2 Métricas de qualidade manuais	29
3.3 Tipologias de Erros: características mais fortes e limitações	32
4. Metodologia	37
4.1 Objetivo do trabalho	38
4.2 Construção de uma tipologia baseada em dados	38
4.2.1 Anotação do lado do Agente EN	40
4.2.2 Anotação do lado do Utilizador PT-BR	43
4.2.3 <i>Source Typology</i> : classes de erros propostas	47

4.2.3.1 Diretrizes para o processo de anotação	69
4.2.3.1.1 Extensão de segmentos	70
4.2.3.1.2 Casos complicados	75
4.2.3.1.3 Severidades	82
4.2.3.1.4 Árvores de decisão	85
5. Discussão de resultados	88
5.1 <i>PT-BR_EN inbounds</i>	88
5.2 <i>Agent Annotation</i>	95
5.3 <i>Multilingual internal pilot</i>	99
5.3.1 Análise de estruturas neutras	101
5.3.2 Análise de erros críticos	114
5.3.3 Uso indevido da tipologia	116
5.3.4 <i>Inter-Annotator Agreement</i>	120
5.3.4.1 Estudo sobre IAA no piloto interno	122
6. <i>Language eGuides</i>	127
7. Conclusões e trabalho futuro	129
Bibliografia	132
Anexos	143

Abstract

In this study we propose a typology which concerns source errors and linguistic structures that might have an impact on Machine Translation (MT). Although most typologies are built on a bilingual level, the source text (ST) also presents issues that cannot be expected to be resolved by MT. In this study, we were able to test whether or not the quality of the ST has an impact on the target text (TT) quality.

For that purpose, source data was annotated. The data analyzed was both inbound (user-generated content) and outbound (agent) in the context of chat. Through this analysis, it was possible to build a data driven typology. To aid the construction of a new typology, there was also a comparison between multiple typologies, whether they have a bilingual or a monolingual focus. This allowed us to see what could be applied to a monolingual typology and what was missing. With the annotation results, it was possible to build a new typology — Source Typology.

To assist future annotators, we provided annotation guidelines with a listing of all the issue types, an explanation of the different span types, the severities to be used and the tricky cases that might occur during the annotation process.

In order to test the reliability of the typology, three different case studies of an internal pilot were conducted. Each case study had a different goal and took into account different language pairs. By testing the Source Typology, we could see its effectiveness and reliability and what should be improved.

In the end, we demonstrated that the quality of the ST can actually have an impact on the TT quality, where, at times, minor errors on the source would become or originate critical errors on the target. The typology is now being applied at Unbabel.

Keywords: Source Typology; Machine Translation; Annotation; User-generated content; Customer Support

Resumo

Neste trabalho propõe-se uma tipologia do texto de partida (do inglês, *Source Typology*) que considera erros no texto de partida (TP) e estruturas linguísticas que têm impacto na tradução automática (TA). Embora a maioria das tipologias seja construída tendo em conta um nível bilíngue, o TP também apresenta problemas que não conseguem ser previstos pela TA. Neste trabalho, foi possível testar se a qualidade do TP tem ou não impacto na qualidade do texto de chegada (TC) e como aferir objetivamente esse mesmo impacto.

Inicialmente, foi efetuada uma comparação com diferentes tipologias de anotação de erros, quer estas considerassem um nível bilíngue ou monolíngue (*e.g.*, *TAUS MQM-DQF Typology*, *MQM Top-Level* e *SCATE MT error taxonomy*, tipologias que serão apresentadas na *Secção 2.4*). Esta comparação possibilitou verificar as semelhanças e diferenças entre si e também quais as classes de erros previamente utilizadas.

De forma a ter mais informações sobre este tema, foi realizada uma análise de dados do TP. Os dados foram analisados em contexto do conteúdo de *chat* e produzidos por utilizadores e agentes. Esta análise foi realizada através do processo de anotação. Este processo permite a identificação e categorização de erros e difere conforme as diretrizes apresentadas. Nesta primeira fase, o processo de anotação foi efetuado na plataforma *Annotation Tool* com a Tipologia de Erros da Unbabel. Uma vez que esta tipologia foi construída num contexto bilíngue, verificaram-se quais os erros que também sucediam no TP.

Além disso, foi possível averiguar, nesta análise, quais eram os erros mais comuns no TP e examinar as diferenças entre um utilizador e um agente. A linguagem de *chat* é bastante específica, trazendo consigo simultaneamente as características da escrita e do diálogo. Enquanto o utilizador tem uma linguagem menos cuidada, algo que dá origem a diferentes tipos de erros, o agente tem de seguir um guião com soluções pré-definidas, atendendo sempre a restrições de tempo. Para além destes restringimentos, os agentes ainda têm de lidar com o facto de, na sua maioria, não serem nativos da língua inglesa, aquela que lhes é requerida no apoio ao cliente, e de ter condições de vida precárias.

Esta análise foi efetuada através de uma das métricas manuais de qualidade mais amplamente utilizada na área da TA — *Multidimensional Quality Metric* (MQM) — proposta no projeto *QTLaunchPad* (2014), financiado pela União Europeia. Assim, os resultados do

processo de anotação foram convertidos de modo quantificável, para aferir a qualidade do TP. Através desta análise, foi possível criar uma tipologia baseada em dados.

Com os resultados desta análise, foi possível produzir uma nova tipologia — a *Source Typology*. Para auxiliar futuros anotadores desta tipologia, foram fornecidas diretrizes para o processo de anotação com a listagem de todas as classes de erros (incluindo as novas adições), esclarecimentos quanto aos tipos de segmentos conforme a anotação pretendida, as severidades utilizadas e os casos complicados que podem surgir durante o processo de anotação. De forma a clarificar esta última secção, também foram fornecidas duas árvores de decisão, uma delas a assistir na classificação de erros ou de estruturas linguísticas e outra a assistir na escolha da severidade adequada.

De modo a comprovar a fiabilidade da tipologia, foi realizado um piloto com três estudos distintos, com um total de 26855 palavras, 2802 erros e 239 estruturas linguísticas (representadas na severidade ‘Neutra’ — associadas a marcadores discursivos, disfluências, emojis, etc., mecanismos característicos do discurso oral) anotados. Cada um dos estudos realizados no piloto abrangeu diferentes objetivos e teve em conta distintos pares de línguas. Em todos os estudos realizou-se uma análise para verificar se os erros encontrados no TP tinham sido originados ou transferidos para o TC e se as estruturas linguísticas com a severidade ‘Neutra’ tiveram ou não algum impacto nos sistemas de TA.

O primeiro estudo, *PT-BR_EN inbound*s, focou-se em PT-BR_EN e considerou textos produzidos por utilizadores. Este estudo foi realizado tendo em conta diferentes clientes da Unbabel. Neste estudo a língua de partida (LP) utilizada foi o português do Brasil e a língua de chegada (LC) foi o inglês. O valor de MQM no TP foi elevado (72.26), pois os erros mais frequentes eram erros de tipografia, ou seja, de baixa severidade. Contudo, ao comparar com o valor de MQM no TC, houve uma grande disparidade. No TC houve muitos erros críticos, algo que não seria de esperar, dada a qualidade do TP. Esta discrepância implicou uma análise mais aprofundada. Desta análise, verificou-se que 34 erros presentes no TP tinham sido transferidos para o TC, 29 erros no TP deram origem a outros erros no TC e houve 9 estruturas neutras que tiveram impacto no TC. Ao examinar diferentes exemplos, observou-se que grande parte dos erros de baixa severidade e as 9 estruturas neutras no TP resultaram em erros críticos no TC.

O segundo estudo, *Agent Annotation*, concentrou-se em textos em inglês produzidos por agentes da área de apoio ao cliente. É importante referir que o inglês não é “nativo”. Ao

contrário do primeiro estudo, este derivou apenas de um cliente, uma vez que os dados dos agentes são dependentes dos clientes específicos e de guiões fornecidos por cada cliente em particular. Neste estudo foram utilizadas duas línguas, o inglês como LP e o francês como LC. Ao contrário do primeiro estudo, o valor de MQM do TC foi mais elevado do que o valor resultante do TP. Porém, também foi realizada a mesma análise neste estudo. 59 erros encontrados no TP foram transferidos para o TC e 40 erros no TP originaram novos erros no TC. Uma grande diferença entre o primeiro e segundo estudo foi de nenhuma estrutura neutra no TP ter tido impacto no TC.

O último estudo, *Multilingual internal pilot*, foi o mais extenso de todos por incluir várias línguas e vários anotadores, tendo em conta tanto o lado do utilizador como o do agente. Relativamente aos estudos prévios, este estudo foi realizado numa escala bem mais alargada. As línguas anotadas neste estudo foram: holandês, italiano, espanhol europeu, português do Brasil, romeno, polaco, alemão e inglês. Os valores de MQM em cada língua diferem de acordo com as diferenças entre línguas e os erros encontrados. Observou-se, nesta análise, que o número de erros foi superior ao número de segmentos, o que significa que, por média, cada segmento apresentava mais do que um erro. Neste estudo, as estruturas neutras com impacto no TC foram divididas por classes e não por línguas devido à extensão de erros. Conjuntamente, também foram apresentadas as suas formas corretas nas LC. O mesmo processo foi realizado para os erros críticos encontrados no TP. Ao longo da análise, também se verificou que algumas classes de erros não foram anotadas de forma correta ou que não foram anotadas quando eram necessárias. Este fenómeno permitiu logo verificar a eficiência da tipologia e das suas diretrizes. Desse modo, são apresentados os casos em que essas situações surgiram e as razões por detrás do sucedido. Para uma análise mais completa, também foi investigado se estes casos tiveram algum impacto no TC. Das 44 estruturas neutras que não foram anotadas no TP, 10 delas tiveram, de facto, impacto no TC.

Ao testar a *Source Typology*, foi permitido ratificar a sua eficiência e a fiabilidade e o que deve ser melhorado. A eficácia da tipologia foi avaliada através do *Inter-annotator Agreement* (IAA), uma metodologia que permite identificar ambiguidades e falhas que resultaram do processo de anotação. O IAA possibilita averiguar se houve ou não concordância entre os anotadores, como também a concordância que os anotadores tiveram consigo mesmos. Outra particularidade do IAA é verificar se os anotadores das mesmas línguas têm a mesma noção de extensão de um erro ou estrutura linguística. Instruções quanto

a este t3pico foram explicitadas nas diretrizes, mas ainda pode haver d3vidas sobre este processo de segmenta33o de erros. Assim, surge uma oportunidade para melhorar essa sec33o nas diretrizes.

Por fim, atrav3s destes estudos foi demonstrado que a qualidade do TP tem, de facto, impacto na qualidade do TC, em que, por vezes, erros m3nimos encontrados no TP se tornam ou originam erros cr3ticos no TC. Estes estudos tamb3m permitiram perceber quais os erros cometidos pelos utilizadores e os agentes e a diferen3a entre eles e, ao mesmo tempo, validar a tipologia, que est3 em produ33o na Unbabel.

Palavras-chave: *Source Typology*; Tradu33o Autom3tica; Anota33o; Utilizador; Servi3o de Apoio ao Cliente

1. Introduction

This dissertation was performed in the context of the Master's degree on Translation of the School of Arts and Humanities of the University of Lisbon. To this end, an internship was carried out at Unbabel. Unbabel is a company that merges machine translation with human post-edition. Its focus is on Customer Support.

One particular aspect that tends to be ignored in the field of Machine Translation is the quality of the source text. Usually, it is not even in question to doubt it. As translators and readers, we assume it will be of high quality, so we try to achieve the same with our translation. Of course, that is not always the case. The same way we question our own translations and the ones performed by MT, that is what should also be done with the ST. A demonstration of this matter is that performing an in-depth data analysis from the source is a novel project at Unbabel. Moreover, the source data analyzed in this context is human-generated and in the context of chat language, which is mostly considered a mixture of written and spoken language, justifying its tendency for errors. Being a company focused on Customer Support, there are two sides to consider: inbound (user) and outbound (agent). Agents are call center employees and users are the clients that ask for support. In both cases, we have different situations. Usually, agents are non-native speakers of English and users, although natives of the language they are typing, have different fluency levels. With user-generated content in chat data, very lively and spontaneous conversations occur and these conversations often have an impact on the MT process. By also focusing on the agents' side, it is possible to gather information on why some errors occur taking into consideration their background of not being English native speakers in their majority of cases and not being provided with the best working conditions. So, to better understand both sides of source data both the inbound and outbound sources will be analyzed. If a translation needs to be reviewed before being displayed to the client, then the source should also be reviewed in order to understand its impact on the translation's quality.

Analyzing source data will allow for a broader sense of the impact that the source's quality will have on the performance of machine translation quality. This will also benefit the training of MT models by checking which errors are the most common and how they can be captured and resolved by the engines. Another aspect that brings attention to this is that

editors at Unbabel have felt uncomfortable or refused to work with a problematic source text and have reported it to the company, so this project will also help to improve their experience.

In the scope of this dissertation, it is essential to build a data driven typology in order to understand which errors can be considered at a monolingual level. For that purpose, a thorough research will be conducted about monolingual and bilingual typologies, especially the former. Given the context of Customer Support, it is also important to examine the particularities of the chat language and of the dynamics of chat conversation (inbound and outbound). Taking all this information into consideration, the annotation process with the Unbabel Error Typology, which concerns a bilingual level, will then allow us to verify, at first hand, which issue types were used in the ST and which ones were missing. After establishing a typology with annotation guidelines and decision trees, it is crucial to test it in an internal pilot with multiple languages and annotators. Thereafter, we can see the reliability of the typology and what needs to be clarified and improved and, more importantly, whether or not the errors and linguistic structures found in ST have an impact on TT.

In *Section 2.*, all Unbabel translation pipelines and quality processes will be explained. In *Section 3.*, the state of the art of MT will be described, highlighting translation quality evaluation metrics and processes and the current typologies used as a reference. In *Section 4.*, the methodology used in this dissertation will be explained. In *Section 5.*, the results of the three case studies of an internal pilot will be presented and discussed. In *Section 6.*, the production of language guides will be illustrated. And finally, in *Section 7.*, the conclusions and intentions for future work will be displayed.

2. Host characterization

Unbabel is a software company that merges human translation with artificial intelligence. It was founded in 2013 by Vasco Pedro, João Graça, Bruno Silva, Hugo Silva, and Sofia Pessanha. At the moment, Unbabel is based in San Francisco, California, Pittsburgh, Lisbon, London and Berlin. The Portuguese startup is currently working with 30 languages, combining multiple language pairs and bringing with it a great variety of employees and customers. Unbabel works with bilinguals, professional translators, and highly skilled linguists. Its focus is the translation of Customer Support content, where three content types are available. These content types will be further explained in the following sections.

2.1 Content types

Unbabel translates Customer Service content, namely tickets (emails from users of a specific customer), chat, and frequently asked questions (FAQs). There are some differences between the three different content types translated at Unbabel, as we will describe.

In the Customer Service jargon, tickets are support email threads concerning customers' doubts or complaints. Chat is the most challenging content translated at Unbabel, since it is used for having the same flow as a normal conversation would occur, but exclusively with text with demanding time expectations. One of the advantages of chat is being live customer support and Unbabel engines are able to translate this support in real-time. Both tickets and chat are one-on-one communication. FAQs are to agilize the customers' self-service content by being available in several languages. FAQs use one-to-many communication, since this content type is mainly used on a client's website. Since FAQs are going to be more available to different people as a front page of a customer, its quality is of great importance, therefore the demand on quality is the priority, not so much the delivery time of the translated material.

2.2 Unbabel translation pipeline

As previously discussed, the content types differ substantially, so in this section, we will be describing mostly the translation pipeline of each content type and how different or similar their flows are.

The client submits their order through CRMs¹ or platforms provided by Unbabel. Then, the company enables different integrations according to each client's needs. These platforms are very important because it is through them that the data to be translated is received and where, then at the end, that the translation is delivered back to the client. Recently, Unbabel developed a Customer Portal in order to scale up the orders, allowing their clients to have access to their subscription, usage data, and translation quality.

2.2.1 Tickets translation pipeline

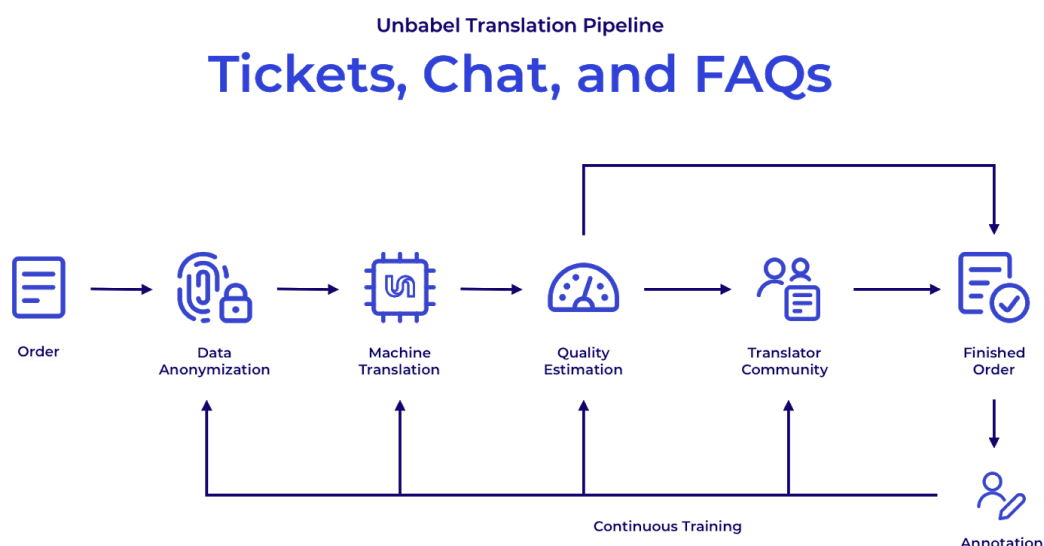


Figure 1. Tickets translation pipeline

As shown in *Figure 1*, before the machine translation, there is a very important step: data anonymization, as in Customer Support content is crucial to maintain the privacy and security of the clients according to the GDPR². This sensitive data is Personal Identifiable

¹ CRM stands for Customer Relationship Management which “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/>)

² GDPR (or General Data Protection Regulation) is a directive of the European Parliament which has “the principles of, and rules on the protection of natural persons with regard to the processing of their personal data

Information (PII), which is usually emitted through Named Entities. Named entities can be names, phone numbers, addresses, currencies, and so on. All PIIs are anonymized by Unbabel’s proprietary system, Named Entity Recognition (NER), and only after that module is applied the texts are automatically translated. The Machine Translation step is performed with Neural Machine Translation (NMT). Afterward, there is the quality estimation step (QE) which automatically detects the quality of the translation. If the QE system evaluates a translation above a certain threshold, the translated content immediately goes to the client. This is called QE skip, because the other steps are skipped due to being unnecessary and time-consuming. However, if the QE score is considered below the conservative threshold, the machine-translated content is delivered to the post-editors’ community. The post-editors’ community is going to edit and correct the errors in the machine-translated text, so it can be finally delivered to the client. The final step is the annotation process, which allows the labeling of errors and the improvement of the MT systems.

2.2.2 Chat translation pipeline

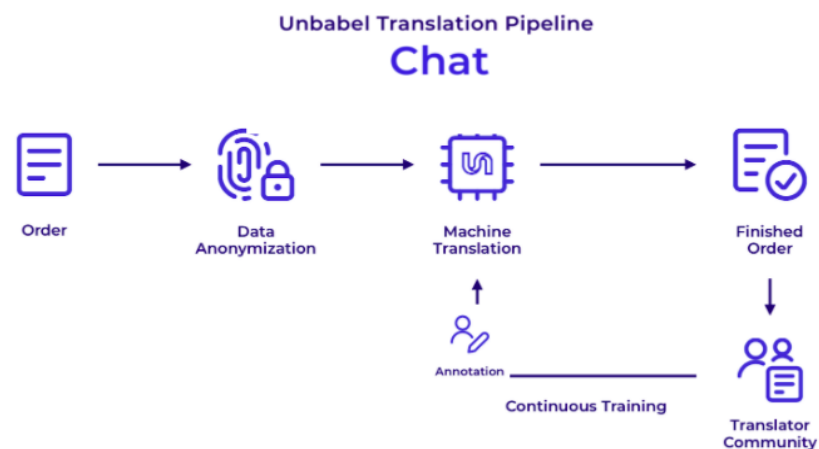


Figure 2. Chat translation pipeline

The translation pipeline for chat is slightly different than the one for tickets. Until the MT module, every step is the same. The main difference is that for chat content there is no QE step or human translation required, as it is illustrated in *Figure 2*. This means that after the content is machine translated, it is directly delivered to the client. The reason for this is

should, whatever their nationality or residence, respect their fundamental rights and freedoms, in particular their right to the protection of personal data.”
<https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN>)

the immediacy that is implied in chat. While tickets are expected to be delivered in a few hours, chat is performed during real-time. The norm for chat is to be delivered in seconds, so if the response's translation delays the whole process this will cost not only the client but also Unbabel. Chat models are trained with post-edited data, which is only used for this purpose.

2.2.3 FAQs translation pipeline

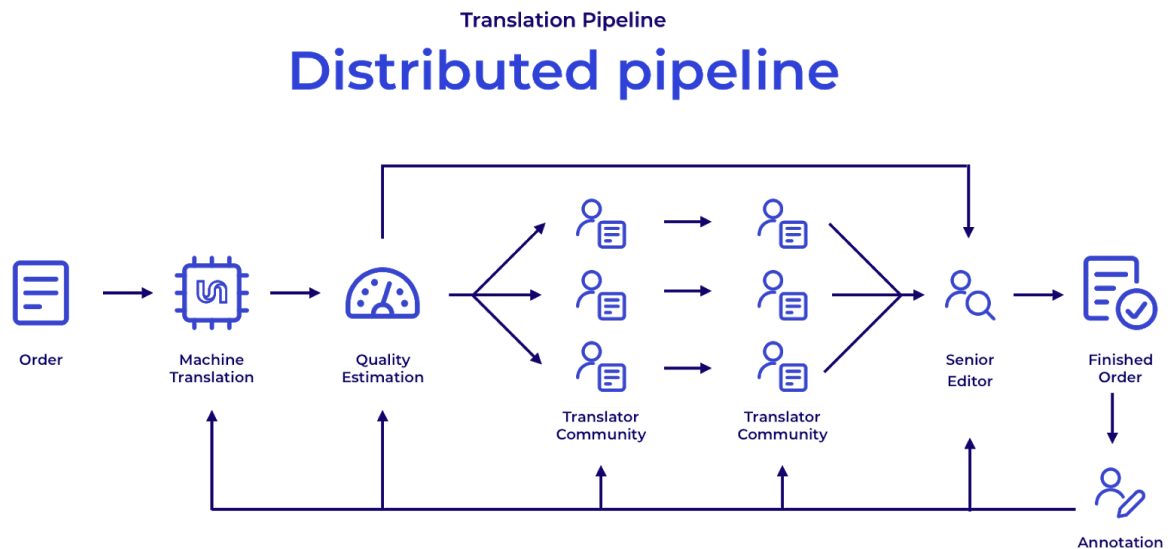


Figure 3. FAQs translation pipeline

The FAQs pipeline is almost the same as the pipeline for tickets, except it does not include QE, and the post-editors' community step is more developed, as demonstrated in *Figure 3*. Since the quality of FAQs is extremely important, as stated previously, the translation is revised by editors and also by a Senior Editor, who proofreads the translation as a whole. Only then the product is delivered to the client.

2.3 Unbabel Quality processes

In order to ensure translation quality, Unbabel performs multiple processes, whether it is QE, Edition, and Annotation. As explained previously, QE is a quality estimation system that predicts the quality of a machine translation output by doing a comparison between the source and target text or source and post-edited text. By achieving a certain threshold, other steps in the translation pipelines are skipped. The edition of machine-translated text is performed by translators, linguists and terminologists. Besides editing, these professionals

also curate linguistic resources requested by the clients, such as glossaries and translation memories (TMs). The Annotation process is usually the last step in all translation pipelines. This process allows us to see MT errors so that these can be resolved before the training phase. Due to its relevance, the annotation process will be further explained in the following section.

2.3.1 Human Evaluation

There are several methods of human evaluation, such as Direct Assessment (DA) or Multidimensional Quality Metrics (MQM)³. We will mainly focus on the latter due to its frequent usage in both industry and research. MQM is a quality assessment framework that allows a personalized customization of quality metrics. This assessment is possible through the process of annotation.

Error annotation is an evaluation process used to identify translation errors and categorize them according to a certain annotation error typology. However, the categorization of errors, even if it was of the same data, can be performed in various ways depending on the criteria used in its evaluation (Lüdeling & Hirschmann, 2015). Its evaluation differences are usually stated in the annotation guidelines provided. Annotation guidelines provide all the information that an annotator needs to know before starting the annotation process. As Burchardt & Lommel (2014) pointed out, the guidelines are not only for training purposes, but also for reference during the annotation process. Guidelines encompass categories and issue types to be used to define an error, if there are any rules on the span selection of an error, what severities will be used to quantify an error and how they will be computed, and also if there is any ambiguity that might come with the typology or taxonomy presented. Besides this, it is important to highlight that annotation is a manual evaluation process, which means that it can also be very subjective. As Lüdeling & Hirschmann (2015) explained, there is not a universal truth in annotation because it implies the interpretation of its reader. That is why they present two different interpretations of an error — a grammar error and a usage error. By having explicit grammar rules in each language, it is easier to identify and explain these kinds of errors. An usage error depends more on interpretation, focusing on “pragmatics, information structure, register, etc.” (Lüdeling & Hirschmann, 2015:140). From

³ <http://www.qt21.eu/mqm-definition/definition-2015-12-30.html>

time to time, while annotating, it is impossible to infer what the data provider was trying to convey. For those cases, first the annotator tries to understand its context, if it is not clear, then the annotator only has one option, which is a reconstruction of the segment or, as Lüdeling & Hirschmann (2015) named it, a target hypothesis. In their paper, they provide an example that describes this accurately.

(1a): “She must saved money.”

In this example, we can observe that there is an error in the sentence concerning the verb tense. Without its context, we can only make some target hypothesis that would have been possible, such as “She must have saved money.” or “She must save money.”. This example specifically could be disambiguated by having the context, making it impossible to select any target hypothesis. With this, it is possible to see that although we can detect a grammar error, “the identification of appropriateness errors needs linguistic and extra-linguistic context.” (Lüdeling & Hirschmann, 2015:141).

In order to be able to annotate, there are some aspects that need to be clarified. Usually the data is separated into tokens, which essentially is the minimal unit to be annotated. Then, according to its guidelines, there are rules of the selection span of an error and how many tokens should be annotated in different circumstances. It is also important to know if there are any limits on the number of annotated tokens because in some cases there are tokens that can contain more than one error. For that to happen, we need to categorize the errors. In order to achieve this, a list of issue types is provided. An issue type “describes the type of the error within a given error annotation scheme.” (Lüdeling & Hirschmann, 2015:146) and it can describe a grammatical, lexical or semantic error. Usually the categorization of errors is displayed in a hierarchical order according to its frequency or relevance. One aspect that is key for annotation is consistency. This will allow us to see how useful and effective the annotation guidelines and everything explained in them, from the error categorization to the selection span, actually are. There are two ways to evaluate the annotation process — a gold standard annotation and Inter Annotator Agreement (IAA) (Lüdeling & Hirschmann, 2015). With the former, there needs to be a correctly established annotation of the data, and with the latter, several annotators are provided equally with guidelines, then assess the same dataset and finally the segments where they agreed on and

the frequency of their agreement is examined. The IAA will be explained in further detail in *Section 5.3.4*. Through this process, it is possible to verify the reliability of the error categorization and whether or not it is clear. This is a continuous process where the guidelines are readjusted until there is consistency in the annotation process at last.

As stated previously, in the scope of MQM, annotation is the process where translation errors are annotated according to a specific typology. After the errors are annotated, they are also rated according to a specified severity. We will mainly focus on the Unbabel Error Typology and explain it broadly. This typology is an adaptation of the one provided by MQM, where some subset of issue types are used.

Currently, the Unbabel Error Typology has three coarse categories — *Accuracy*, *Fluency* and *Style*. *Accuracy* concerns whenever “The target text does not accurately reflect the source text”, *Fluency* concerns “Issues that affect the reading and the comprehension of the text” and finally, *Style* concerns whenever “The text has stylistic problems”. This typology has a total of 16 daughter issue types, 30 granddaughter issue types and 10 grand-granddaughter issue types. The severities used are *Critical*, *Major* and *Minor*. A minor error is when it “doesn’t lead to a loss of meaning and it doesn’t confuse or mislead the user” but it can decrease the stylistic quality or the fluency of the text. A major error is when “it misleads the user; the change of meaning results in the improper use of the product/service; it appears in an important part of the text content”. A critical one is when the meaning of the original text is changed and carries health, safety, legal or financial implications or if it damages the company’s reputation or misrepresents the functionality of the product/service. Each of these severities has a corresponding value, with *Critical* being 10 points, *Major* 5 points and *Minor* 1 point. All annotations at Unbabel are performed on an in-house platform — Annotation Tool. These are computed to calculate the MQM score. With an MQM score, we cannot only quantify the quality of a translation, but also profusely analyze the errors found. For further information on MQM, please refer to *Section 3.2.2*.

Through annotation, Unbabel can evaluate not only the machine translation and QE outputs, but also the post-edition, and therefore give feedback to their editors and give feedback to AI teams. The annotation process is performed by highly experienced translators and linguists, who first have training and are provided with the Unbabel Annotation Guidelines, and then have regular feedback in order to improve their skills.

2.4 Evaluation process: evaluating the post-editors' community

Unbabel has many communities involving editors, senior editors, and experts, such as terminologists, evaluators, language consultants, and annotators. The community of post-editors allows a better quality of translations by checking the accuracy of the machine translated text and improving the MT engines on future translations.

In order to ensure the quality of its translations, Unbabel mainly focuses on the evaluation of the work delivered and on the editors involved. These evaluations are regularly made with criteria concerning *Style* (register used), *Fluency* (grammatical and orthographic errors), and *Accuracy* (meaning). These editors start with a testing phase where they get acquainted with language and usability guidelines, then training, and finally paid tasks. Their quality assessment is through the rating from 1 to 5 stars, as presented below in *Figure 4*:



Figure 4. Editor's quality assessment

These evaluations are made by professional translators and linguists. If the editor is still in the training phase, the evaluation will determine the possibility of becoming a Paid Editor. A positive evaluation will mean gaining the status of a Paid Editor and a negative evaluation will mean that the editor will not have access to paid translations. The same happens in a paid editor evaluation. If the evaluation is positive, then their paid status is maintained and if the evaluation is negative, the editor will be demoted to Trainee Editor.

As exemplified, Unbabel takes into account how its different content types require their own translation pipeline. Each pipeline has its own order and steps working accordingly with their specificity. In order to maintain the quality of these content types, the company also ensures the quality of the professionals that work on them by having a regular structured evaluation process.

3. State of the art

In this section, it is presented the historical progression of Machine Translation (MT), the significance of quality metrics, automatic and manual, and a comparison between multiple typologies that were vital for the research and creation of a new source annotation typology. Technology has found its ways to merge in every aspect of our lives and Translation is no exception. Initially with translation memories, being only used as support tools, and then with MT, the impact of technologies has been pervasive, as this section will show.

3.1 Machine Translation

Nowadays, MT is very much embedded in Human Translation (HT) and its fast-growing development allowed new possibilities for HT. The early beginnings of MT started in the 1930s in France, with Georges Artsrouni, and with his creation of a *Cerveau Mécanique* ("Mechanical Brain"), and Petr Trojanskij in Russia, with the first automatic translator. Trojanskij's device had three steps:

“... a monolingual human operator would parse a source text using a universal scheme that could capture all possible grammatical functions of words. The operator would then locate source-text words, one by one, in the part of the machine that acted as a translation dictionary, and add the relevant grammatical code for the current use of that word. The machine would output the ‘equivalent’ word in the target language, along with the grammatical code, information that would, in a third step, be used by a monolingual target-language speaker, to create a morphological correct target text.” (Kenny, 2018:429-430)

In 1949 Warren Weaver, of the Rockefeller Foundation, drafted a Memorandum where he suggested that the use of electronic computers would solve the problem that translation presents, due to the diversity of languages and cultures (Kenny, 2018). Instead of adopting a word-for-word translation, Weaver proposed to have an immediate context of a word by adding an essential number of words at the right and left of such word so that there was not any place for ambiguities. Weaver was also in favor of working on a universal language. Without doubt, this sparked criticism and skepticism against machine translation.

However, Weaver's Memorandum was also an object of motivation for MT research in the early 1950s. One example of this is the case of a working MT system at Georgetown University in 1954. Léon Dostert, a language scholar known for the great historical impact on the Nuremberg Trials on account of his interpretation system, was invited to participate in this project alongside IBM (International Business Machines Corporation). Both showed a Russian-to-English translation system. Despite some criticism, this system was mostly seen as an omen of the great success of MT, mainly because it seemed as being used in the near future. Due to the uneasy times during the Cold War, both the US and the Soviet Union started funding MT research groups. Yet, these countries had different aims. The US mainly focused on Russian-to-English translation, resembling the Georgetown University project, while the Soviet Union was thinking more broadly by taking into account several other languages, such as French, German, Chinese, Czech and Bulgarian (Bar-Hillel, 1960).

Even though the 1950s put MT in the spotlight, it faded with time resulting in the interest of other areas. In 1960, Yehoshua Bar-Hillel stated that even with Weaver's proposal of adding context of n words at the left and right of an ambiguous word would not erase ambiguity altogether. This could be the case of an idiom of the source language and the risk of it being translated literally, as already analyzed by Bar-Hillel (Bar-Hillel, 1953). Therefore, the linguist justified that only with encyclopedic knowledge would ambiguity no longer be a problem in MT. However, Bar-Hillel doubted that the machine could have that sort of knowledge, so the quality of the translation would always be compromised. Still, what really stagnated the interest and enthusiasm over MT was in 1966 through the Report from the Automatic Language Processing Advisory Committee (ALPAC). There were multiple concerns on MT research. The hegemony of the English language in scientific literature became more evident over the years, hence not justifying the use of translation. Its main concerns were quality, speed, and cost (Hutchins, 1996). There were no means or metrics to evaluate the quality of the translations, the translation process was still very slow, and despite the advances of MT, it still needed human intervention in the shape of post-editing. Essentially, the ALPAC Report stated that MT research was a field that needed to be more developed and the current funding for it was not reasonable. As a consequence of this, MT research in the US started to decrease.

Even so, the research began to increase in other parts of the world, such as Canada and the Commission of the European Communities (CEC). With the Official Languages Act

in 1969, Canada declared its bilingualism as a nation. This act resulted in the constant use of both English and French. MT research was mostly focused on one area — weather forecasts, resulting in an MT system called Météo system. This became a very advantageous decision because weather forecasts were linguistically simple with a limited “range of vocabulary and grammatical structures” (Kenny, 2018:432), required fast translations for being momentary, were replaced quickly by a new forecast, and were also considered dull to HT.

Around the same time, the European Union (EU), CEC then, was also setting foot in MT research by using the Systran MT system. The number of EU Member States was increasing, resulting in several language pairs and thus the vital need for translation. The Systran MT system was used until 2010. Later on, the EU started to invest and use in-house MT systems.

In the 80s, MT interest began to grow in North America and in Japan, where the computer market started to expand. This gave origin to the creation of commercial systems. Simultaneously, post-editing became a more common practice in order to prevent problems in MT. Another practice that also took place was pre-edition. Pre-edition was done in source texts which implied using rules on grammar, such as the choice of vocabulary. One case in particular of trying to improve MT was the Automated Language Processing System (ALPS). Although this system was initially developed to “translate religious tracts simultaneously into several target languages” (Sugden, 1985:403), it was used for assisting the translator, word processing, dictionary lookup, and the correction of translation errors (Sugden, 1985). The ALPS system also resolved ambiguities found in the source texts. However, there were a few disadvantages about the system. It was not prepared for idiomatic speech and in order to have high-quality translations, it needed post-editing. This need for post-edition and the software itself were considered too expensive at the time, so in consequence, it was not embraced by the MT community.

By the end of the 20th century, a new MT system emerged — Statistical Machine Translation (SMT). An SMT is a system where the translation is achieved through bitext or parallel data and the frequency of sentences. A bitext is the source text aligned with the corresponding HT. Additionally, SMT has a language model for target languages based on extensive monolingual corpora of those target languages and performs a statistical analysis of n-grams, also known as phrase-based. Despite having precise translations, they lacked fluency and consistency, a lot of words were omitted and if a language was morphologically

rich its translation might not be as good (Koehn, 2020). During this period, the IBM approach was what stood out the most. Its learning through bitext without the use of linguistic knowledge amazed all the MT community. This system allowed more availability of electronic bitext and an increase of computer power and data storage.

Finally, in the 21st century, the interest in some languages began to grow. Arabic and Mandarin became a necessity for the US, due to political and economic conflicts. The EU recognized many languages as its official languages, highlighting the case of Irish because this allowed a greater availability of that language. In a short period of time, technology found its way to grow substantially, leading to a broader online linguistic diversity and a bigger visibility and availability of MT systems. It was recently that a new data-based system emerged — Neural Machine Translation (NMT). NMT uses neural networks, including machine learning as a part of Artificial Intelligence (AI). Its learning consists of large quantities of training data, something that only became possible as a consequence of the Internet's never-ending growth. NMT translations, contrasting with SMT, might not be the most precise but they are more fluent (Koehn, 2020). Although this system can handle languages morphologically richer, it still consists of words or segments' omission and also additions that have no relation with the source language. Currently, NMT is considered the state-of-the-art in the MT field.

3.2 Translation Quality Evaluation

A translation's quality is vital because it defines the value of the translation produced. One of the consequences of MT was the growth of translation quality metrics. Nowadays, there are two types of quality metrics — automatic metrics and manual metrics. Before having these quality metrics, a professional translator/linguist would review both source and target texts, usually someone bilingual, to determine a translation's quality. However, this practice had a major flaw. Without any predefined tools, most evaluations were informal and the translations and reviews provided would disperse from one to another, thus creating an inconsistency that made clients doubt the final product (Moorkens *et al.*, 2018). So, quality metrics were a fresh innovation to quality evaluation. Throughout the years, there have been concerns about the quality evaluation methods, which one should be used or how they should be improved. Like any other field, quality evaluation has been developed over the years and

quality metrics have been valuable to the field of translation by providing a diversity of metrics with different functions.

3.2.1 Automatic quality metrics

Throughout the years, automatic quality metrics have been developed. These metrics provide a value illustrated through an algorithm concerning its quality, however they do not provide any information on the errors present in a translation. Its results allow developers of MT automatic metrics to observe the effect of daily changes to their systems (Papineni *et al.*, 2002). Automatic metrics need to be sensitive to small differences, consistent with their scores, applicable to different languages and domains and also needs to be fast. Some metrics stand out from the majority, such as BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit ORdering), and very recently BERTScore, BLEURT (Bilingual Evaluation Understudy with Representations from Transformers) and COMET (Crosslingual Optimized Metric for Evaluation of Translation), Unbabel's proprietary metric.

These automatic metrics evaluate the quality of MT output compared to reference human translation, that should be created by a professional linguist. BLEU considers multiple reference translations, allowing to use a different word choice for a translation of the same source word. BLEU is also a precision-based metric that computes the precision for different types of n-grams and combines the precision scores of the different n-grams into one single score. One of the reasons for this system to stand out was due to the multiplicative brevity penalty, which meant that a high scoring translation must match the reference translations according to length, word choice and word order (Papineni *et al.*, 2002). The BLEU metric extends from 0 to 1, where 0 is a bad translation and 1 is a perfect translation. It is very unusual for translations to get 1 as a score, unless they are identical to a reference translation.

METEOR was developed later, taking into account the fragilities of BLEU and thus building a more efficient system. This system was based on “an explicit word-to-word matching between MT output being evaluated and one or more reference translations” (Banerjee & Lavie, 2005:2). One of the great advantages of METEOR is that it can be extended to include more advanced matching strategies. METEOR works by aligning the correlation between the metric score and human judgments of translation quality and it

estimates a score for the matching using these features: unigram-precision, unigram-recall and a measure of fragmentation (how well or badly ordered are the words in the MT output) (Banerjee & Lavie, 2005). METEOR's matching also covers identical words, words "that are a simple morphological variants of each other" (Banerjee & Lavie, 2005:2) and words that are each others' synonyms.

One of the latest automatic metrics is BERTScore. This metric computes "token similarity using contextual embedding" (Zhang *et al.*, 2019:1). Firstly, it is given a reference translation and a candidate translation, then the contextual embeddings are used to represent the tokens, and finally, the matching is calculated with a cosine similarity (Zhang *et al.*, 2019). With a main model, BERT (Bidirectional Encoder Representations from Transformers), "which is an unsupervised technique that learns contextualized representations of sequences of text" (Sellam *et al.*, 2020:7882), the input text is tokenized into a sequence of word pieces, "where unknown words are split into several commonly observed sequences of characters" (Zhang *et al.*, 2019:4). The similarity measure between reference and candidate considers isolated tokens, while the contextual embeddings contain the information of the sentence. The BERTScore matches each token from the reference to a token from the candidate "to compute recall", and the opposite happens "to compute precision" (Zhang *et al.*, 2019:4). However, this metric also presents disadvantages. It depends too much on the quality of the embedding coming from the underlying models and its scores are usually very high, where even unrelated sentences can have degrees of similarity.

Another recent automatic metric is BLEURT, a learned automatic metric based on BERT developed by the Google Team. This metric stood out due to a pre-training scheme that uses synthetic examples in order to help the model to generalize (Sellam *et al.*, 2020). BLEURT uses a reference sentence and a prediction sentence which will then be estimated with the training dataset that will predict the human rating. This model was trained for English by extracting grammatically diverse sentences from Wikipedia in order to be evaluated by different generalized systems. As a result of being a learned metric, "BLEURT can model human assessment with superior accuracy" (Sellam *et al.*, 2020:7888).

Finally, Unbabel's proprietary metric COMET, which is a "neural framework for training multilingual machine translation evaluation models" (Rei *et al.*, 2020:1). These evaluation models gather information from both the source-language input and the

target-language reference translation, which needs to be of high quality, and an hypothesis to predict the MT quality more precisely. The reason for using the source input was because the source helped the models to learn accurate predictions and improved the COMET ranking. This metric stands out due to two architectures — the Estimator model and the Translation Ranking model. The distinction between them is their training purpose. The Estimator model calculates a quality score, while the Translation Ranking model shortens the distance between “a “better” hypothesis and both its corresponding reference and its original source.” (Rei *et al.*, 2020:2). COMET also has novel attributes, such as not punishing the phrasing of a reference and being sensitive to gender, whether there is agreement or not, and to formality. Unlike the previous metrics, COMET has a very high correlation between agreement and human judgment, due to being trained with DA, Human-targeted Translation Error Rate (HTER), and MQM data.

Besides these metrics, there is also Quality Estimation (QE), which is an automatic quality estimation system. QE does a comparison with the source text, the target text and the post-editing performed in the target text. QE predicts automatically the quality for a machine translated output without having access to any reference translations (Specia *et al.*, 2018b as cited in Kepler *et al.*, 2019). QE allows the assessment of the MT output, highlighting whether or not it needs human post-edition (Kepler *et al.*, 2019). QE can be performed on word-level or sentence-level. Word-level QE enables a more fine-grained estimation by classifying each word as ‘bad’ or ‘good’ in terms of translation quality or if any word from the source text was omitted. Whereas sentence-level QE predicts the quality of the whole sentence and provides insight if the MT output needs to be edited. Unbabel has also created its own framework concerning translation QE — OpenKiwi (Kepler *et al.*, 2019). OpenKiwi is an open source framework that supports “training and testing of word-level and sentence-level quality estimation systems” (Kepler *et al.*, 2019:117).

3.2.2 Manual quality metrics

Manual quality metrics provide a quality score and allow a very fine-grained categorization of translation errors. The starting point of these metrics was when Language Service Providers (LSPs) started to move toward a systematic quality evaluation. The tools used initially were spreadsheets and they allowed the reviewers to “count numbers of errors

to generate overall quality scores, usually represented as a percentage, with 100% indicating no errors” (Moorkens *et al.*, 2018:111). Besides this, the reviewers would also assign weights concerning different severities. However, these initial tools also had problems by having a single reviewer to compute the final scores and only them knowing why it presented that score, and by being impossible to verify if the final scores were generated with the clients’ requirements in mind.

In the 1990s, the standardization of systematic quality evaluation began with two projects — SAE J2450 and LISA QA Model. The former project was developed by SAE international and presented “a simple scorecard-style for automotive documentation” (Moorkens *et al.*, 2018:112), featuring six error types and two severity levels. The latter project was developed by the now-ceased Localization Industry Standards Association (LISA), “which released it as a spreadsheet and later as stand-alone software.” (Moorkens *et al.*, 2018:112). Opposite to SAE J2450, the LISA QA Model was more specific and featured 18 or 21 categories, some issues only concerning localisation into East Asian languages, and three severity levels. This project was designed for two content types “(documentation and software user interface)” (Moorkens *et al.*, 2018:112). Despite promoting transparency, both projects fell short of its promises due to its restrictions. The Inter-annotator Agreement (IAA) was low, having disagreement on the error severity and on the errors themselves. The standardization of issue types did not merge well with models that had specific scenarios or text types in mind. As a result, the LISA QA Model was constantly altered in order to adapt to the specificities presented and the SAE J2450 had to state that “it was intended only for automotive service manuals and that other content types would require their own metrics.” (Moorkens *et al.*, 2018:112-113). In the late 2000s, LISA was to release a new project, Globalization Metrics Exchange — Quality (GMX-Q), to complement the limitations of the LISA QA Model; however, due to LISA’s shutdown in 2011 this was not possible. As a consequence of the decline of these projects, two groups started focusing on translation quality assessment — the Translation Automation User Society (TAUS) with its Dynamic Quality Framework (DQF) and the EU-funded QTLaunchPad project with its Multidimensional Quality Metrics (MQM).

The TAUS DQF system "addresses a variety of approaches to quality assessment, including those aimed specifically at MT, such as measuring post-editor productivity, adequacy/fluency evaluation, readability (...) and crowdsourced evaluation” (Moorkens *et al.*,

2018:123). This system stood out by taking a different direction, TAUS reached out to LSPs and clients and asked which were the best practices for this quality evaluation. This allowed TAUS to create a simple typology mainly focused on “the needs of its localisation-oriented members” (Moorkens *et al.*, 2018:123). Just as the LISA QA Model, the first release of this typology was in Excel format incorporating instructions about where to enter error counts for each of the categories with the four severity levels available.

MQM is a flexible quality assessment framework that allows users to define custom metrics for quality evaluation. This system was developed to “address the shortcomings of previous quality evaluation” (Lommel *et al.*, 2014:458) and adopted some ideas developed in the GMX-Q project. One of the aims of MQM was to have a system that could evaluate translation’s quality as objectively as possible in a short amount of time. Another aim is that “MQM is intended to be language neutral and therefore applicable to any language pair” (Lommel *et al.*, 2014:459). Instead of having a one-size-fits all model, its flexibility allows unlimited customization for its users. Nevertheless, the QTLaunchPad project is still used as reference. MQM can be used for several purposes, whether it is for the evaluation of commercially-produced translations or for the evaluation of academic translations.

While MQM was firstly designed to assess target texts’ quality, later began to also be used in the assessment of source texts and the impact that it might have in the target text. One of the great advantages of MQM is that it can be used to evaluate any type of text, whether it is machine or human translated. For quality assessment, there are several factors that need to be considered, such as the categorization for quality issues and the scoring mechanism. In order to define and organize issue types, there needs to be a hierarchy, with a tree-like structure, in an error typology where parent issues are followed or not by children and grandchildren issues. The more ramified the hierarchy, the more specific the issue is. In case the categorization of issues becomes extensive, decision trees were designed to help to differentiate the multiple issues presented. Initially, the MQM typology had 104 issue types and currently it has 182 issue types⁴. Concerning the scoring mechanism, this is also a very flexible tool and it is performed through severities, usually resulting in different scoring and the use of different severities according to the user. However, firstly, severity “refers to the nature of the error itself and its effects on usability of the translation.” (Moorkens *et al.*, 2018:120). The severities used in this typology are *Critical*, *Major*, *Minor* and *Null*. Other

⁴ <http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html>

severities have also come up, such as *Kudos* in the current harmonized TAUS MQM-DQF error typology. According to the procedure adopted, each severity corresponds to penalty points that will then be used in the calculation of the total MQM score. In order to compute the total MQM score, each error is collected and it is then multiplied “by its severity value and its weight to generate penalty points.” (Moorkens *et al.*, 2018:121). There are no set thresholds of what means a good and bad score in MQM, only that the perfect score is 100. However, in Sanchez-Torron & Koehn (2016:21) 95% was considered the minimum quality acceptance level of a professional post-edition. This became widely accepted in the academic community and for that reason this score will also be used as a reference in this dissertation.

Although these two systems clashed in the beginning, currently the entities that developed them are in contact and are trying to work together for their common purpose and, despite their differences, the two systems complement each other. While MQM “provides a way to describe arbitrary metrics in a standardized fashion” (Lommel *et al.*, 2014:461), DQF provides “guidance on interpreting quality evaluations for specific scenarios.” (Lommel *et al.*, 2014:461). By joining these systems together, a new restructure was essential. This resulted in reorganizing the dimensions of MQM in order to match DQF’s top-level categories, adding a new severity level (*Null*) for issues that needed to be marked without any penalty falling on them, adopting MQM issue names, expanding its categorization of issues, and adding new dimensions (*Internationalisation* and *Verity*) (Moorkens *et al.*, 2018). With these changes, a new harmonized DQF-MQM error typology was implemented. This typology “has become the preferred method to implement MQM” (Moorkens *et al.*, 2018:125) and its “inclusion in DQF has helped raise the profile of the MQM approach to TQA.” (Moorkens *et al.*, 2018:125).

Having quality assessment metrics, whether it is manual or automatic, is essential to MT. Without it, the evaluation process would be slower, less efficient and more expensive. These systems allow a more objective evaluation and a more fine-grained categorization of issue types found in source and target texts.

3.3 Error Typologies: strengths and limitations

Currently, there are several error typologies on a bilingual level that take into account the relationship between the source and the target text. However, the number of typologies

concerning only source errors is very scarce. According to the MQM typology, there are multiple errors that are not specific to a bilingual level, they are just as frequent in the source text. In the listing of their issue types⁵, there is a table for each issue type with its definition, an example and whether or not that issue type applies exclusively to the target or source text or both of them. Here is an example of that with the *Agreement* issue type:

Agreement

ID	agreement
Definition	Two or more words do not agree with respect to case, number, person, or other grammatical features
MQM Core?	no
Automatable?	yes
Parent	word-form
Children	none
Applies to	source and target
Example(s)	<ul style="list-style-type: none"> • A text reads “They was expecting a report.”

Figure 5. MQM typology issue type example

As seen in *Figure 5*, the *Agreement* issue type can be applied to the source and target because an issue concerning agreement can affect the source text the same way as it can affect the target text.

One typology that took into account errors on a monolingual and bilingual level separately was the SCATE MT error taxonomy⁶ (Tezcan *et al.*, 2017). Once acquainted with this typology and the Unbabel Error Typology, an initial comparison was made. The Unbabel Error Typology is more fine-grained, as it is more specific when it comes to grammar errors. However, being fine-grained also presents a difficulty. Due to its specificity, the typology becomes rather extensive, and that way more difficult to learn it at first and apprehend all the issue types present in it, making the annotation process slower than it has to be and the inter annotator agreement could be lower. As presented below in *Figure 6*, the SCATE typology, on a bilingual level (highlighted in purple), also seems very complete. However, on one hand, on a monolingual level (highlighted in orange), there are certain issue types that do not exclusively belong to it. For example, errors such as *Orthography* (which includes

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

⁶

https://www.researchgate.net/publication/323626831_SCATE_taxonomy_and_corpus_of_machine_translation_errors

Punctuation and Capitalization) and *Word Order*. These errors can also be seen in a target text simultaneously. With this, there is an assumption that MT does not make mistakes which in itself would be incorrect since it is something occurring. They usually come from NMT systems and are denominated as “hallucinations”.

On the other hand, we have errors that are only considered on a bilingual level, such as *Addition*, *Omission* and *Part-Of-Speech*. This is assuming that the source text is being produced perfectly, which cannot be the case if it is user-generated.

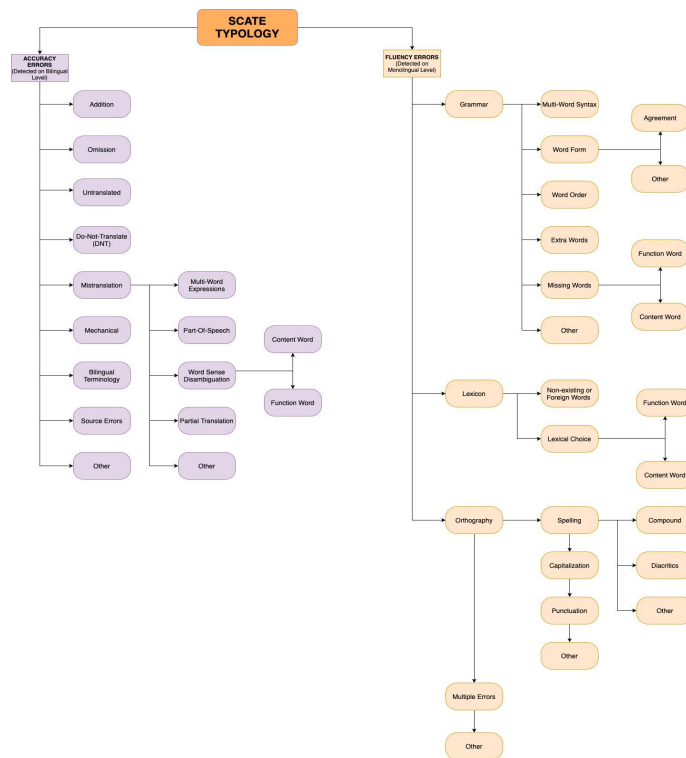


Figure 6. The SCATE MT Error Taxonomy

The Unbabel Error Typology was built on a bilingual level, so it is only focused on translation errors. However, we will demonstrate how we used this typology for the development of a monolingual one in *Section 4.2*.

TAUS⁷, an independent organization whose purpose is to help and develop the translation industry by sharing knowledge and data, also has its own typology — TAUS MQM-DQF Typology (Translation specific). Although this typology is very fine-grained, it can also be vague. For example, with the issue types *Awkward* and *Other*. The *Awkward* issue

⁷ <https://www.taus.net/>

type concerns whenever the text is written in an awkward style, and this definition in itself is very ambiguous because what might seem awkward to one person, might seem normal to another and it implies too much subjectivity. The *Other* category is used whenever there are other issues related to the text. It seems strange that a typology so fine-grained needs this category. If the goal is to have a typology that covers as many issues as possible that might occur in a text, then having this category seems superficial.

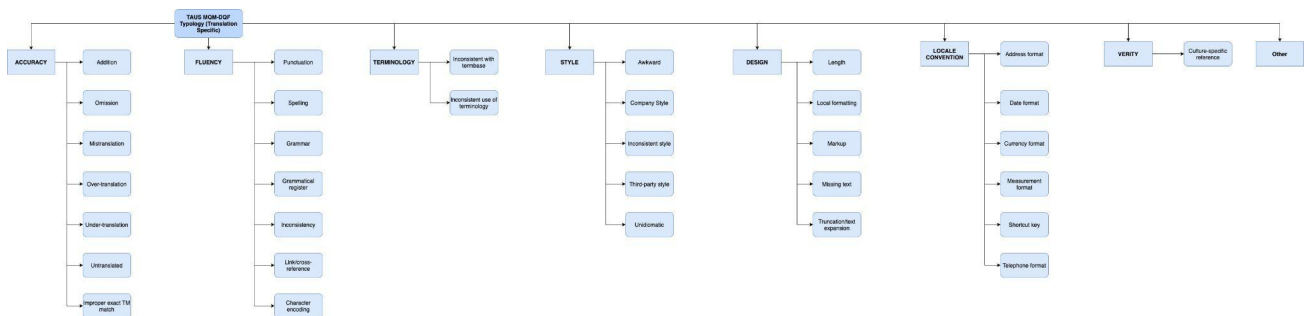


Figure 7. TAUS MQM-DQF Typology (Translation Specific)

TAUS has also set out severity levels. The standard severities that typologies usually adopt for their error categories were included: *Critical* (“Errors that may carry health, safety, legal or financial implications, violate geopolitical usage guidelines, damage the organization's reputation, cause the application to crash or negatively modify/misrepresent the functionality of a product or service, or which could be seen as offensive.”); *Major* (“Errors that may confuse or mislead the user or hinder proper use of the product/service due to significant change in meaning or because errors appear in a visible or important part of the content.”); and *Minor* (“Errors that don't lead to loss of meaning and wouldn't confuse or mislead the user but would be noticed, would decrease stylistic quality, fluency or clarity, or would make the content less appealing.”)⁸. Yet, TAUS also presents *Neutral* and *Kudos* severities. The *Neutral* severity is “used to log additional information, problems or changes to be made that don't count as errors”⁹ and can be a preference of style. This severity can also be found in MQM Core. The *Kudos* severity is used for praise. All severities, except for *Kudos*, were used as reference to the new typology presented in this work.

There is a typology very similar to the TAUS MQM-DFQ Typology (Translation specific) called MQM Top Level. This typology was created by Arle Lommel (2019) and

⁸ <https://www.taus.net/qt21-project#harmonized-error-typology>

⁹ <https://www.taus.net/qt21-project#harmonized-error-typology>

posted in W3C¹⁰. It only differs from one addition in the category *Terminology* with the issue type *Wrong Term*, which is when “The wrong term appears (and is incorrect, regardless of any guidance from a termbase)”¹¹. This can be verified in *Figure 8*.

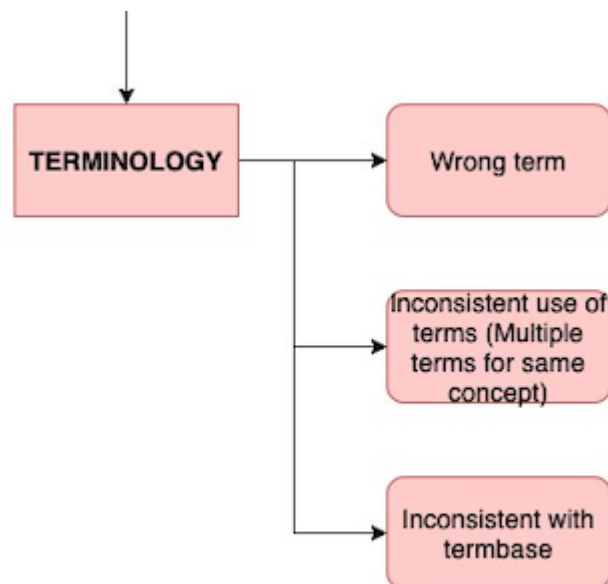


Figure 8. MQM Top Level typology

In conclusion, despite not having a typology exclusively related to source errors, there are multiple typologies that take into consideration that some errors are not solely related to the target text. The SCATE MT error taxonomy, the TAUS MQM-DQF Typology (Translation specific) and MQM Top Level typologies also take into account errors found in both source and target text. However, these typologies are very fine-grained. Although this could be perceived as an advantage in some cases, this also presents difficulties to its annotators and it will eventually decrease the inter annotator agreement. Another issue that the latter typologies presented is having issue types too subjective and vague, such as *Awkward* and *Other*. These issue types can then be used incorrectly. On the other hand, the TAUS MQM-DQF Typology (Translation specific) and MQM Core present an interesting severity, which is the *Neutral* severity. It might be the case that there are some linguistic structures or technological defaults that might have an impact on the text, even though they are not necessarily an error.

¹⁰ “The World Wide Web Consortium (W3C) is an international community where members, organizations, a full-time staff, and the public work together to develop Web standards.” (<https://www.w3.org/>)

¹¹ <https://www.w3.org/community/mqmcg/mqm-top-level-2019-04-11/>

4. Methodology

Although the focus on ST errors in the MT field is very recent, its interest has been growing exponentially over the years. So, creating a typology concerning only errors found in the ST is still an innovative practice in MT. Besides having a monolingual typology, it was also important to see how the errors annotated actually had an impact on the TT. Instead of only focusing on the research available, it was vital to also experiment with real data. Since the beginning of the process, it was decided to build a data driven typology because only then we could have concrete proofs of what could be found in the ST.

Therefore, firstly, we needed to annotate the ST, this will be further explained in *Section 4.2*. In order to reach the production of a monolingual typology, there needed to be a previous annotation effort. As previously stated, currently there is only one typology available at Unbabel, denominated “Unbabel Error Typology”, which mainly focuses on translation related issues. So, the first step was to start annotating the ST with the typology available. Despite not using some issue types, this effort allowed us to verify that some errors that are usually considered as only translation specific errors were actually found in the ST. Apart from that, by already using a typology we could verify if there were any issue types missing. This process will be explained in more detail in *Sections 4.2.1* and *4.2.2* where we annotate the agent and user sides respectively.

After that annotation effort, it was possible to gather all the issue types that would be of use in a monolingual context. Although there were some issue types in common with the Unbabel Error Typology and the MQM typology, it was always our intention to create our own definitions in order to fit in the typology proposed, but maintaining the alignment with the core typologies. In regards to the new additions, it was important to provide substantial information on them and present examples. The proposed typology can be found in *Section 4.2.3* where we describe the creation of the typology.

Once we have established a typology, denominated Source Typology, it was time to create complementary tools to aid the annotation process during the testing phase. So, to complement the Source Typology, we provided annotation guidelines. These guidelines will have all the information available about the Source Typology and eventually help annotators during the annotation process. Within the guidelines we provided the definitions of the issue types with its corresponding examples, explanations on how to annotate concerning the span

types, the tricky cases that might come up if there is some confusion with the issue types, a description of the severities used in this typology and its respective decision trees. All of this information on the annotation guidelines and its content can be found from *Sections 4.2.3.1 to 4.2.3.1.4*.

4.1 Project's objective

In recent years, the focus on the ST in the MT field has been increasing. Traditionally, its main interest was on the relationship between the source and target text. In various contexts, we can assume that the ST is being produced in a correct way. However, given that Unbabel only works with Customer Support content, the source presented is not the most reliable. The source is always human-generated, coming from agents and users, and the content here analyzed, chat, is never corrected or revised due to the immediacy implied in customer service dialogues.

The purpose of this project is to create a typology concerning errors from the ST and look into the impact that it might have on the target text. Aside from being a recent interest in the MT field, it is also a brand new project at Unbabel. Since the beginning of this project it was decided to have real-life examples of source text errors. Therefore, the decided approach was having a data driven typology. This could only be achieved through error annotation.

Even the translation of Customer Support content has not been fully explored, so it was important to consider both sides of a chat conversation instead of just one of them. The level of proficiency, the environment and the emotions involved have a tremendous impact on the errors found in the ST. This analysis also allowed us to have a greater understanding about the origin of source errors. In addition to source errors, we also took into notice linguistic structures that, although are not necessarily errors, have an impact on the MT.

4.2 Building a data driven typology

In order to create a monolingual typology, it was important to first annotate the source. Instead of only relying on previous research or personal assumptions, this would allow us to have concrete proof of what sort of errors are found in the ST. Given its shortage of information in the MT field, annotating directly the ST is an essential step. The method

involved using the Unbabel Error Typology, which is a bilingual typology. Only then, it would be possible to see which issue types would be most frequent in a source text and confirm that many errors are not exclusive to translation per se and can be found also in monolingual contexts. This annotation process also allowed us to understand what other errors should be included that were not present in the Unbabel Error Typology. The results of these annotations will show the frequency of each error, which will complement a monolingual typology by keeping or removing issue types.

The source data annotated was from both agent and final client, which will be named as user, on chat. Before explaining the process of source annotation and its results, there will be a brief introduction to Chat language. This type of language is very specific. Firstly, it is very recent and thanks to the Internet, which is growing by the minute, is always changing and ever-growing. Therefore, being so difficult to explain and capture in its entirety. Secondly, Chat language is an electronic discourse, and as Jonsson (1997) as cited in Nasr *et al.* (2016:175) has described, an “electronic discourse is neither writing nor speech, but rather written speech or spoken writing, or something unique”. As previously stated in *Section 2.2.2*, once the chat content is machine translated, it is directly delivered to the client without any QE step or human translation. So, having this in mind, the variety and number of issue types will be far greater than it would be expected in tickets, for instance.

The reason for annotating the source from two different subjects, agent and user, is because they are very distinct. While agents have templates and terminologies to follow according to their company’s policy (resulting in a more controlled interaction), users are free to write what they want and express more emotions, especially unpleasant ones such as frustration. The source language from the agent was in English and the source language from the user was in Portuguese. The data annotated covered clients from different areas in order to have a greater diversity of errors instead of having the same repeatedly. These areas covered technology, fitness, e-commerce, gaming, courier services, and clothing. *Table 1* shows the results of this preliminary annotation:

	Agent	User
Annotated conversations	170	179
Annotated words	31,440	12,862
MQM	80.32	27.29

Table 1. Source annotation results

This annotation effort was performed by the author of this dissertation in order to have a better understanding of the errors found during the annotation process. Despite annotating more conversations in the user source, the number of words annotated in the agent source is higher due to templates with troubleshooting instructions that agents have to follow with their clients. In the section below, we will show in more detail the errors found in both the agent and user annotations.

4.2.1 Agent annotation EN

It is important to highlight that we decided to show these results in the methodology of our project because this is just a testing phase before starting a pilot (to be presented in *Section 5*) with our proposed typology. Before showing the results in more detail, it is important to take into account some aspects. There are several factors that influenced the number of errors found in agent generated content. To have a better insight about the agents' side and the work environment they are in, there was a meeting with the Unbabel VP of Global Alliances, Edmund Ovington, who has a lot of experience concerning call centers. All centers are mostly based in India and the Philippines and agents are usually found in stressful circumstances, such as struggling in night shifts, sleeping in their cars because the workplace is too far from their home, being in their first job ever, or in the first week of working there. Most agents do not have English as their native language and this can be shown through some errors that could only happen to non-natives of a language, such as *Addition* and *Omission*. Agents also have to answer clients in a short turnaround time in order to close their cases and some errors, such as typos, can result from that lack of response time given to them. Besides these constraints, agents also have to deal with emotional clients and

have different dialogues than they were expecting. The latter aspect will be discussed in more detail. So, knowing the circumstances of the agents helps to better understand the reasons behind some errors that were found. *Figure 9* will display the total number of errors and the issue types used.

Errors distribution (%) - Agent (Unbabel error typology)

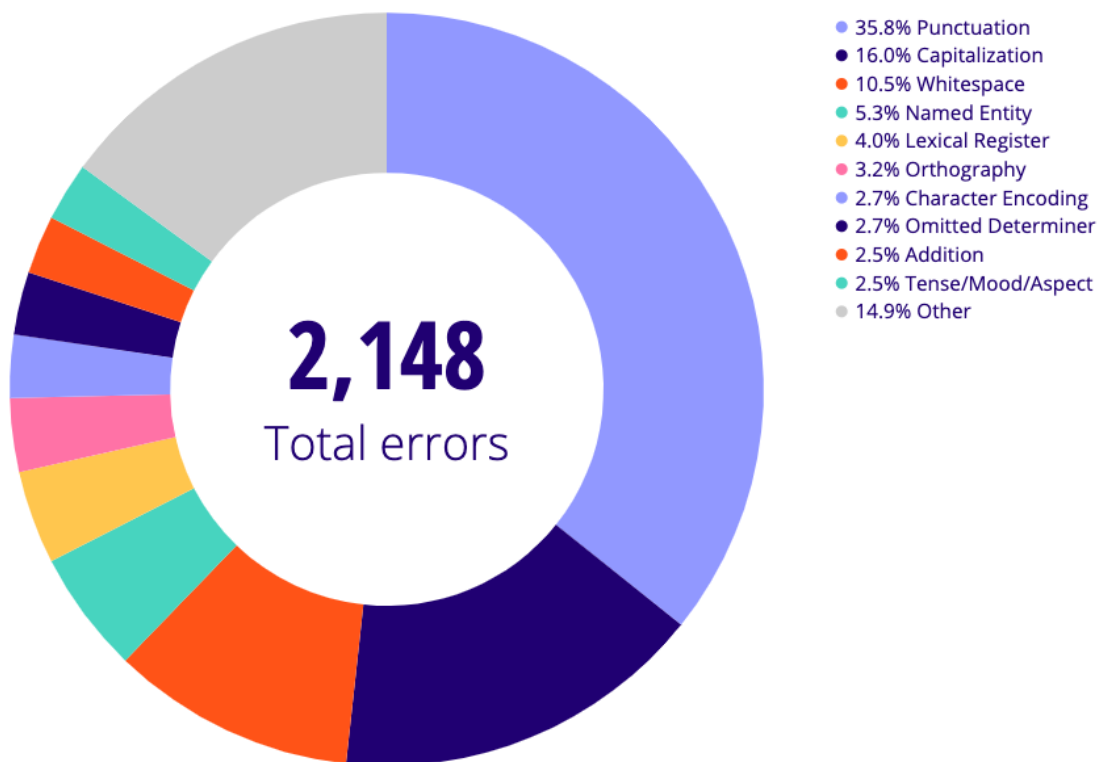


Figure 9. Total number of errors found in the agent annotation EN

Figure 9 shows that most common issue types used in the agent annotation were *Punctuation* (35.8%), *Capitalization* (16.0%), *Whitespace* (10.5%), *Named Entity* (5.3%), *Lexical Register* (4.0%), *Orthography* (3.2%), *Character Encoding* (2.7%), *Omitted Determiner* (2.7%), *Addition* (2.5%), and *Tense/Mood/Aspect* (2.5%). The definition of the most used issue types are according to the Unbabel Error Typology, as previously stated is a bilingual typology for MT, therefore, mentions of the relationship between source and target text:

- *Punctuation*: “Punctuation is used incorrectly or is missing, or one of a pair of quotes, brackets or punctuation — e.g., “ ”, ‘ ’, (), [], { }, ¿, ? , or ¡ ! — is missing from the target text.”;

- *Capitalization*: “Wrong use of capital letters or absence of capital letters.”;
- *Whitespace*: “Whitespace is used incorrectly (there is an extra or missing whitespace).”;
- *Named Entity*: “Named entity tag must be applied in two situations: (1) when names, places, locations or other named entities do not match between source and target; (2) when any other type of error (capitalization, orthography, transliteration, etc.) falls upon a named entity.”;
- *Lexical Register*: “The text uses lexical expressions that are not compliant with the register required for that specific text.”;
- *Orthography*: “Words spelled incorrectly.”;
- *Character Encoding*: “Characters are garbled due to incorrect application of encoding.”;
- *Omitted Determiner*: “A determiner is missing in the target text.”;
- *Addition*: “The target text includes a unit not present in the source.”;
- *Tense/Mood/Aspect*: “A verbal form displays the wrong tense, mood, or aspect.”.

Errors distribution (%) - Agent (Unbabel error typology)

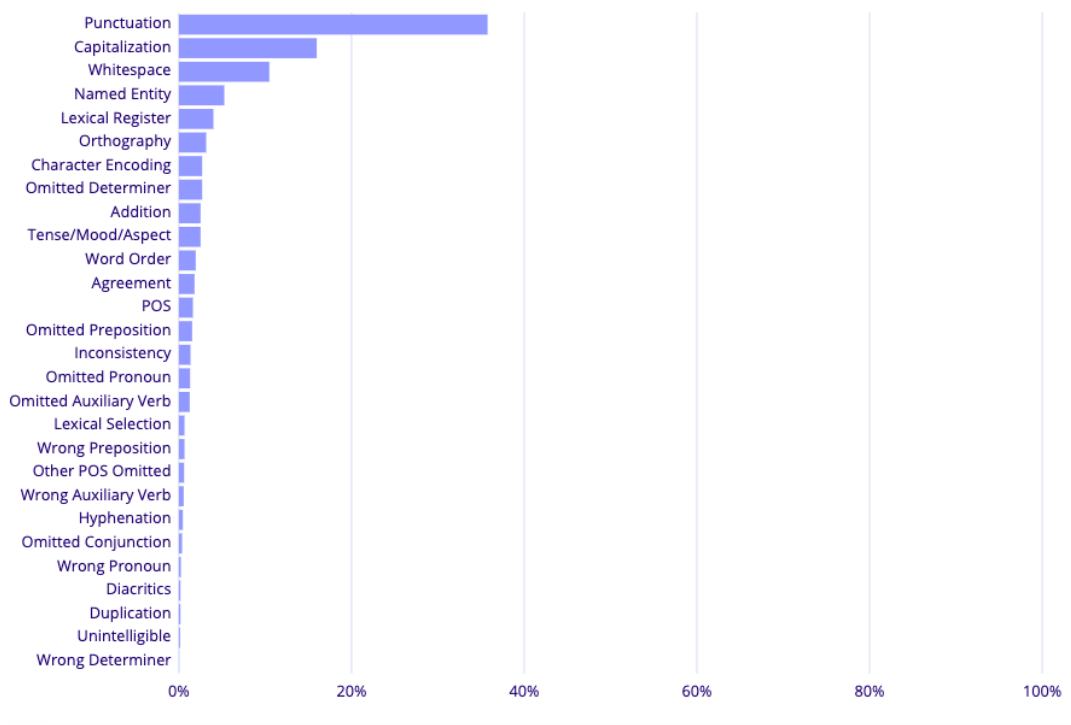


Figure 10. Total of issue types used in the agent annotation EN

While other issue types were less than 2.5%, there were still several of them. In *Figure 10*, the other issue types, such as *Word Order* and *Omitted Preposition*, further confirm that English is not the native language of most agents and that they transfer the rules of their own native languages through omissions or using the wrong prepositions.

4.2.2 User annotation PT-BR

With user-generated content, the context is completely different and the errors found change significantly. One of the reasons for that is due to the content type chosen for this experiment. With chat, people tend to be less careful with punctuation or orthography due to time requirements or even emotional states. In this case, the data annotated was in European Portuguese and Brazilian Portuguese. A previous analysis, performed when testing the Unbabel Typology, showed that Brazilian Portuguese had a great disparity between the data. So, it was decided to also analyze this variety to see the differences between both varieties.

Errors distribution (%) - User (Unbabel error typology)

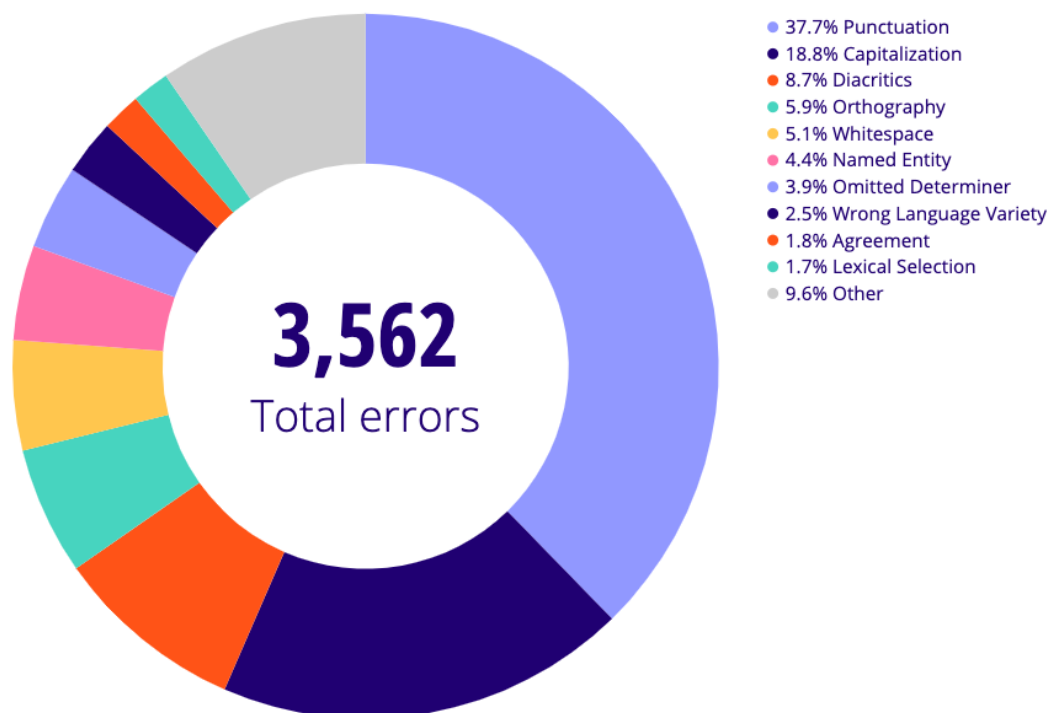


Figure 11. Total number of errors found in the user annotation PT-BR

Figure 11 shows that most common issue types used in the user annotation were *Punctuation* (37.7%), *Capitalization* (18.8%), *Diacritics* (8.7%), *Orthography* (5.9%), *Whitespace* (5.1%), *Named Entity* (4.4%), *Omitted Determiner* (3.9%), *Wrong Language Variety* (2.5%), *Agreement* (1.8%), and *Lexical Selection* (1.7%).

The definitions of the issue types will again be presented according to the Unbabel Error Typology, although not repeating the same errors that were also used in the agent annotation.

- *Diacritics*: “Issues related to the use of diacritics (i.e., any mark placed over, under, or through a letter in some languages, to show that the letter should be pronounced differently). This tag must be applied when the word as a wrong diacritic (another diacritic must have been used), has a diacritic missing or has an extra diacritic.”;
- *Wrong Language Variety*: “The language variety used is not the requested one.”;
- *Agreement*: “Two or more words do not agree with respect to number, gender, person or case”;
- *Lexical Selection*: “The term selected is not correct in context or doesn’t accurately convey the meaning of the original text.”

Other issue types were less than 1.7%, as can be seen in *Figure 12*:

Errors distribution (%) - User (Unbabel error typology)

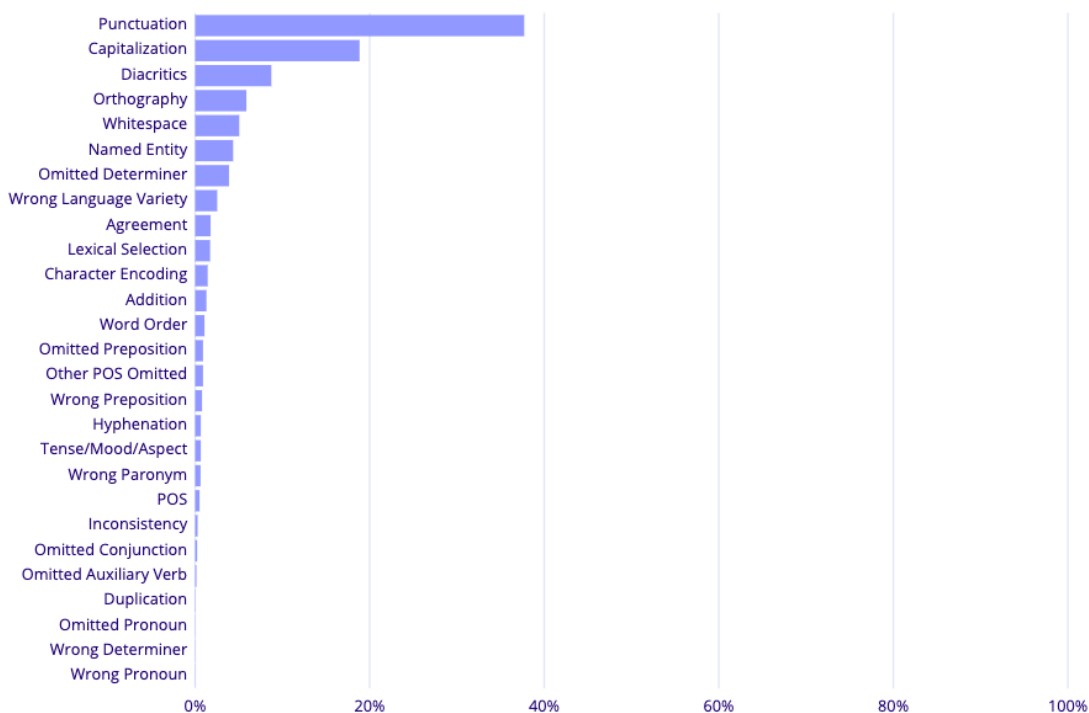


Figure 12. Total of issue types used in the user annotation PT-BR

The number of errors in the user annotation is substantially higher than in the agent annotation. This further shows how native speakers of a language still can make a lot of errors and how problematic a source text can be.

One curious occurrence during this annotation effort was seeing that many of the issue types used in both annotations were *Accuracy* errors. An *Accuracy* error is when “The target text does not accurately reflect the source text, allowing for any differences authorized by specifications.”¹² Since this was source annotation, the high frequency of issue types in this parent issue type seems odd given the MQM definition for *Accuracy*. *Accuracy* is mostly related to the relationship between both source and target text, however the errors were annotated on monolingual level. *Accuracy* in the sense of monolingual data means the target grammar of the speaker was not fully accomplished or, in other words, the uttered message does not fully comply with the intended message. One reason for this is language transfer. Language transfer is whenever linguistic features of one language interfere with another language. This commonly refers to the case of bilinguals, polyglots, or even someone who is

¹² <https://www.taus.net/qt21-project#harmonized-error-typology>

learning a second language besides its native language. In this context, it will be referring to the learning of a second language. “A second language is typically an official or societal dominant language (e.g. English) needed for education, employment and other basic purposes.” (Sinha *et al.*, 2009:117), this is the case of the agents, who live in African and Asian countries. Learning a second language is not an easy task and it can bring many challenges, such as “cognitive constraints and incomplete knowledge of the vocabulary, grammar, and culture” (Goh, 2000; Bloomfield *et al.*, 2011 as cited in Chang & Mishler, 2012:2700). Although language transfer has its positive aspects, its negative effects are what mostly stands out from this process. “The greater the differences between the two languages, the more negative the effects of interference are likely to be.” (Dwinastiti, 2013 as cited in Sirbu, 2015:375). The frequency of errors produced in the second language depends mainly on these differences. While learning a language, “mispronunciation and grammatical errors are the most common types of interference between the mother tongue and the target language” (Marinque, 2013 as cited in Subandowo, 2017:205). Grammatical errors may occur on different levels, whether it is lexis or pragmatics.

Regarding other errors, as can be observed in *Section 4.2.2*, the issue type *Wrong Language Variety* was used. The use of this issue type resulted from different situations. While annotating in European Portuguese, Brazilian Portuguese became more and more frequent. This is not a surprise in any sense since this variety is spoken by a great number of people; however, these varieties have many differences in grammar, spelling, and phonetics. One reason for this is that these two varieties get mixed up by other language speakers. As a native speaker of European Portuguese, it is possible to see how rare it is when a website provides both varieties in their Customer Service. It is rarely translated into the Portuguese language and when that happens, only one of the varieties is offered. Another possible reason for this error is also the amount of Brazilian Portuguese people currently living in Portugal, therefore appearing in the European Portuguese data.

When it came to the data from the user, the error of *Lexical Register* was never annotated since the final client is not aware of the formality or informality of the register of the company they are contacting. And the error *Named Entity* is not seen the same way as it is in the agent data. Sometimes, the users would only have *Orthography* or *Capitalization* problems concerning named entities.

It is possible to see that *Orthography* occurred more often in the user data, besides having more orthographic errors, it was also the case of many abbreviations.

The issue types *Agreement* and any issue type related to *Omission*, such as *Omitted Determiner* or *Omitted Preposition*, mainly appeared in the Brazilian Portuguese data. These are errors that should not be expected when native speakers of a language are writing in their mother tongue, however it can still happen.

4.2.3 Source Typology: proposed issue types

With the annotations' results, it was finally possible to propose a new typology, tailored on evaluating the quality of a source text. Firstly, all the issue types used in both agent and user annotation were listed. According to their high frequency, some issues were considered and included in the new typology. By also making a comparison between several typologies, it was possible to use or combine different categories and issue types that were not included in the Unbabel Error Typology.

The coarse categories in the Source Typology are *Accuracy*, *Fluency*, *Style* and *Design*. Usually, *Accuracy* is specifically thought for translation, so its definition concerns when “The target text does not accurately reflect the source text, allowing for any differences authorized by specifications.”¹³. However, during the annotation process we could see that many sources were not written by native speakers, so we decided to adjust its definition. Although *Accuracy* is usually linked with the correlation between source and target texts, *Accuracy* can also have a different meaning when concerning the learning of a language. In an article provided by TeachingEnglish of the British Council, accuracy refers to “how correct learners' use of the language system is, including their use of grammar, pronunciation and vocabulary.”. So, we kept *Accuracy* in this typology having the definition presented by TeachingEnglish in mind. In the new typology, *Accuracy* concerns the “Mapping between A (actual source written by a user or an agent on-the-fly) and B (intended source). This category is used when the semantic meaning or the conceptualization of an idea is compromised.”. The same was done with the definition of *Fluency*, which addresses “Issues related to the form or content of a text, irrespective as to whether it is a translation or not.”¹⁴. We decided to make

¹³ <http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html#accuracy>

¹⁴ <http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html#fluency>

its definition more source-specific, so currently *Fluency* addresses “Issues that affect the reading and comprehension of the text. Whether or not the text can be read as a native text produced by a native person.”. The definition of *Style* in the typology is the one that is used in the current version of MQM (“The text has stylistic problems.”¹⁵). Although the *Design* category is not present on the Unbabel Typology, it was added because it is a coarse category of MQM whose definition is “There is a problem relating to design aspects (vs. linguistic aspects) of the content.”¹⁶

In this typology, we not only took into account source errors, but also linguistic structures that might have an impact on MT. For these structures, it was decided to annotate them as “neutral”, where no penalty falls on them. Information on these structures can be found in *Section 4.2.3.1.3*. Before presenting the definitions of the issue types, we will first have a distinction between the issue types defined in the typologies provided by MQM and Unbabel and the ones that only concern our proposed typology. Next, we will list all the issue types present in the Source Typology and explain in greater detail their definitions and illustrate them with examples.

SOURCE TYPOLOGY	MQM	UNBABEL ERROR TYPOLOGY
Accuracy	Accuracy (target)	Accuracy (target)
Addition	Addition (target)	Addition (target)
Omission	Omission (target)	Omission (target)
Named Entity	Entity (target)	Named Entity (target)
Lexical Selection	✗	Lexical Selection (target)
Wrong Paronym	✗	Wrong Paronym (target)
Incomplete sentence	Completeness (source)	✗
Fluency	Fluency (source and target)	Fluency (target)
Grammar	Grammar (source and target)	Grammar (target)

¹⁵ <http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html#style>

¹⁶ <http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html#design>

Wrong Function Word	Function words (source and target)	Function Words (target)
Agreement	Agreement (source and target)	Agreement (target)
Tense/Mood/Aspect	Tense/Mood/Aspect (source and target)	Tense/Mood/Aspect (target)
Wrong POS	Part of speech (source and target)	Part of speech (target)
Word Order	Word Order (source and target)	Word Order (source and target)
Typography	Typography (source and target)	Typography (target)
Capitalization	Capitalization (source and target)	Capitalization (target)
Diacritics	Diacritics (source and target)	Diacritics (target)
Hyphenation	Hyphenation (source and target)	Hyphenation (target)
Orthography	Spelling (source and target)	Orthography (target)
Punctuation	Punctuation (source and target)	Punctuation (target)
Whitespace	Whitespace (source and target)	Whitespace (target)
Code Switching	✗	✗
Style	Style (source and target)	Style (target)
Register	Register (source and target)	Grammar/Lexical Register (target)
Wrong Language Variety	✗	Wrong Language Variety (target)
Emoticon	✗	✗
Conversational Marker	✗	✗
Idiomatic	✗	✗

Profanity	Offensive (source and target)	X
Abbreviation	X	X
Design	Design (source and target)	X
Markup	Markup (source and target)	Character Encoding (target)
Numeration	X	X
Segmentation	X	X

Table 2. Common issue types between Source Typology, MQM and Unbabel Error Typology

As it can be observed in Table 2, there are a lot of similarities between the three typologies. Most of the issue types are not only exclusive to the TT, but also to the source. Although, there are some exceptions. While the issue type *Completeness* in the MQM typology concerns only the ST, we have six issue types (*Accuracy*, *Addition*, *Omission*, *Entity/Named Entity*, *Lexical Selection* and *Wrong Paronym*) that only concern the TT. This was no surprise given the previous discussion about *Accuracy*. Another observation worth highlighting is that there are some issue types that are not covered in these two typologies, which consist in the enrichment of the Source Typology. These are the new additions proposed in our typology.

Throughout the research, having non-native speakers, in our particular case mainly agents, was also taken into account. This involved looking into two types of experiments: pre-editing and errors produced by non-native speakers.

Pre-editing is “the process of rewriting the source text (ST) to be translated in order to obtain better translations by MT” (Miyata & Fujita, 2021) and this brings out many advantages. One of them is allowing the learning of a second or foreign language for non-native speakers of said language. In Shei (2002), an interesting experiment took place where pre-editing tasks with an MT system were used for teaching English as a foreign language to Chinese students. With pre-editing, it is possible to create a set of rules that will disseminate a percentage of errors in the MT output. One example of this is a study carried out by Mercader-Alarcón & Sánchez-Martínez (2016) where the creation of pre-editing rules for English improved Spanish MT output by 11%.

To better understand the difficulty that a language may present, it is necessary to study how that particular language translates as a second or foreign language. By observing and analyzing the different kinds of errors that non-native speakers produce, we can fully understand a language. Given that non-native speakers can be found in both sides of customer support, especially in the agents' side, it was also important to learn more about this study.

In the following part, we will present all the issue types of the Source Typology, divided into three sections. Firstly, we will introduce the issue types that our typology has in common with the other two typologies. Secondly, we will present the issue types that were previously considered as only related to the TT. Finally, we will show the new additions of the Source Typology that are not present in any of the other typologies.

We will present the common issue types between the typologies:

Accuracy:

Incomplete sentence: “The sentence is truncated and it’s impossible to infer its meaning. This results in the abandonment of the sentence’s semantic concept and sometimes a refresh of the idea intended. This would account for more than one token missing.”

Despite having a different name in MQM (*Completeness*), this issue type was essential to the Source Typology. Sometimes agents have trouble finishing an idea because they have time restrictions while working in troubleshooting solutions in real time. For users, the scenery changes slightly. While writing online, it is common to encounter disfluencies. Disfluencies “are usually defined around an interruption point, where the sentence flow is interrupted.” (Gilmartin *et al.*, 2017:25). According to Shriberg (2001) as cited in Gilmartin *et al.* (2017), over a third of utterances in natural conversations are disfluencies. In a study performed by Kraichoke (2017), where it was taken into consideration a lot of the errors found among learners of English as Foreign Language (EFL) and English as a Second Language (ESL), it was also found an error very similar to *Incomplete Sentence*, denominated “Sentence Fragment”. Its definition stated by Kraichoke (2017) is:

“A sentence fragment is a group of words that do not form a complete sentence, nor express a complete thought. Sentence fragments typically are portions of sentences, disconnected from the main clause, and often lack a subject or a verb”. (Kraichoke, 2017)

This kind of error would not only have an impact in the source text, but also in the target text.

EN (source):

(6a): “Please try to sign in to the link and try to check if a “Continue” or “Next” button [Ø] [Ø].”

Fluency:

Wrong Function Word: “When a function word is used incorrectly”.

Functions words are “exemplified by prepositions, articles, auxiliary verbs, pronouns, and such-words whose principal role is more syntactic than semantic.” (Smith & Witten, 1993:3). It was decided to go with a more general issue type, instead of having multiple specific issue types, such as *Wrong Preposition* and *Wrong Determiner*. This approach was used so that the number of issue types would be reduced and then the typology would not seem so extensive. Besides this, using a more general issue type will help make the typology easier to learn for the annotators, which will then help increase the IAA. Deciding on more general issue types instead of being too fine-grained would also help the typology to work with different languages that have different language structures. By doing the opposite, we would be risking having very specific issue types, while there would be other ones missing.

When studying the errors produced by non-native speakers, prepositions and articles are given the most attention to. Gamon (2010) explains that the interest in studying articles and prepositions in learners of English is motivated by two reasons: “They are a closed class and they comprise a substantial proportion of learners’ errors”. Concerning articles, Gamon (2010) shows that “The candidates for article choice are *the* and *a/an*, and the choice for prepositions is limited to twelve very frequent prepositions (*in, at, on, for, since, with, to, by, about, from, of, as*) which account for 86.2 % of preposition errors in our learner data.”.

During the analysis process, the author verified that half of the sentences for articles and prepositions were flagged as having at least one error in them.

EN (source):

(7a): “If you have any other questions, please let me know **or** I'll be glad to help.”

(7b): “If you have any **another** questions, please let me know.”

Agreement: “Two or more words that do not agree with respect to number, gender, person, or case.”.

We have added case agreement in order to include as many languages as possible.

PT (source):

(8a): “**A aplicações** estarão disponíveis na próxima semana.”

DE (source):

(8b): “Sie läuft bis zu **den Eingang**.”

Tense/Mood/Aspect: “A verbal form displays the wrong tense, mood, or aspect.”.

It was important to keep this issue type in the typology because this issue occurred in both agent and user data. Agents have a particular difficulty with verb tenses, especially with the past tense in English. Despite most users being native speakers of the language they are writing, they still had a lot of problems with verbs. Usually they would write the way that is orally spoken, which sometimes is not grammatically correct. One study performed by Miyata & Fujita (2021) analyzed how pre-editing could improve NMT and even created a typology with 39 issue types. The language-pairs used in this study were JA-EN, JA-ZH and JA-KO and the NMT engines used were from Google and TextTra. The typology created in that study is provided in *Figure 13*:

ID	Editing operation type	Ja-En		Ja-Zh		Ja-Ko		Total	Expl.	Impl.	Pres.
		G	T	G	T	G	T				
S01	Sentence splitting	1	0	3	3	4	3	14	0	0	14
S02	Structural change	3	5	9	4	4	2	27	8	1	18
S03	Use/disuse of topicalisation	1	7	4	3	1	3	19	5	2	12
S04	Insertion of subject/object	2	1	1	3	5	2	14	14	0	0
S05	Use/disuse of clause-ending noun	3	2	2	2	2	1	12	12	0	0
S06	Change of voice	1	3	0	0	0	0	4	2	0	2
S07	Other structural changes	1	0	2	1	1	0	5	3	0	2
P01	Insertion/deletion of punctuation	19	16	5	12	9	10	71	0	0	71
P02	Use/disuse of chunking marker(s)	6	12	2	1	3	4	28	11	8	9
P03	Phrase reordering	6	4	7	1	9	4	31	0	0	31
P04	Change of modification	1	3	3	0	0	0	7	0	0	7
P05	Change of connective expression	3	18	4	2	10	3	40	24	5	11
P06	Change of parallel expression	3	8	2	8	4	11	36	7	2	27
P07	Change of apposition expression	1	7	2	1	1	4	16	8	4	4
P08	Change of noun/verb phrase	1	3	2	1	3	3	13	9	3	1
P09	Use/disuse of compound noun	1	5	2	2	6	12	28	16	12	0
P10	Use/disuse of affix	4	4	1	2	3	3	17	1	0	16
P11	Change of sahen noun expression	0	1	1	1	2	0	5	1	0	4
P12	Change of formal noun expression	1	2	2	2	2	0	9	4	0	5
P13	Other phrasal changes	0	1	0	1	2	1	5	4	0	1
C01	Use of synonymous words	18	18	19	18	25	20	118	14	10	94
C02	Use/disuse of abbreviation	2	7	2	2	1	7	21	19	2	0
C03	Use/disuse of anaphoric expression	4	4	2	2	1	1	14	10	2	2
C04	Use/disuse of emphatic expression	1	2	2	1	4	1	11	10	1	0
C05	Category indication/suppression	5	3	6	5	4	7	30	29	1	0
C06	Explanatory paraphrase	3	4	1	0	1	1	10	0	0	10
C07	Change of content	22	20	21	9	14	8	94	57	23	14
F01	Change of particle	9	14	4	6	7	7	47	13	5	29
F02	Change of compound particle	8	5	5	2	5	6	31	24	2	5
F03	Change of aspect	1	4	1	0	5	1	12	0	0	12
F04	Change of tense	0	0	1	1	1	1	4	0	0	4
F05	Change of modality	3	1	2	1	3	1	11	5	0	6
F06	Use/disuse of honorific expression	3	1	1	2	2	1	10	0	0	10
O01	Japanese orthographical change	10	16	9	5	9	12	61	12	4	45
O02	Change of half-/full-width character	0	5	3	2	2	4	16	7	1	8
O03	Insertion/deletion/change of symbol	0	2	0	0	0	0	2	0	0	2
O04	Other orthographical change	0	1	0	0	3	0	4	0	0	4
E01	Grammatical errors	0	8	5	2	2	5	22	-	-	-
E02	Content errors	5	0	8	1	1	1	16	-	-	-

Table 5: Constructed typology of editing operations (G: Google, T: TexTra). The first letter of ID indicates the six major categories (S: Structure, P: Phrase, C: Content word, F: Functional word, O: Orthography, E: Errors casually introduced in the ST). The right three columns provide the frequencies for general informational strategies (Expl.: Explication, Impl.: Implication, Pres.: Preservation).

Figure 13. Typology of editing operations (Miyata & Fujita, 2021:7)

It was possible to see that there are several issue types in Miyata & Fujita's typology (2021) common with the proposed typology but, of course, with different denominations, such as *Sentence Splitting*, *Insertion of subject/object*, *Insertion/deletion of punctuation*, *Phrase reordering*, *Use of synonymous words*, *Use/disuse of abbreviation*, *Change of aspect*, *Change of tense*, *Grammatical errors*. We will explain these issue types in their respective definitions below. As it can be observed, *Change of aspect* and *Change of tense* are also included in this study.

EN (source):

(9a): “I already **mark** the order as delivered.”

PT (source):

(9b): “Gostaria de saber o que falta para **pode** fazer chamadas.”

Wrong POS: “A word has the wrong part of speech. The lemma is correct, but the POS is wrong. This is a lexical word where the lemma is always the same, but it is the suffix that creates the POS changes, i.e an adjective instead of a noun.”.

EN (source):

(10a): “I will surely share your feedback with the **concern** team.”

PT (source):

(10b): “Dá para ver que não está a 100%, pois tem alguns pontos falhando **intermitente**.”

Word Order: “The order of the words is incorrect.”.

This issue type was kept the same because it occurred in both agent and user’s data.

PT (source):

(11a): “Entrei no site para fazer a compra e aparece **que o produto não tem**.”

EN (source):

(11b): “I’m sorry, we only have the **black color** in stock.”

Capitalization: “Wrong use of capital letters or absence of capital letters.”.

This is something that is very common in chat language, so it was important to keep this issue type.

EN (source):

(12a): “Sure, **Thank** you.”

PT (source):

(12b): “**bom** dia, Kate.”

Diacritics: “Issues related to the use of diacritics (i.e., any mark placed over, under, or through a letter in some languages, to show that the letter should be pronounced differently). This issue type must be applied when the word has a wrong diacritic (another diacritic must have been used), has a missing diacritic or has an extra diacritic.”

Although English does not use diacritics in its grammar, romance languages, such as Portuguese and Italian, use diacritics in abundance.

PT (source):

(13a): Isso é para todas as **funções** de treino, não **so** corrida.”

Hyphenation: “Misuse of hyphen (the source text is hyphenated incorrectly, has a hyphen missing or has an extra hyphen).”

PT (source):

(14a): Estou tentando desde **quinta feira** cancelar a minha conta.”

EN (source):

(14b): “I request you to **un-install** the application.”

Orthography: “Words spelled incorrectly. This usually is related to typos. This usually results in a non existing word.”

One interesting experiment performed by Stymne *et al.* (2017) analyzed the annotation of texts produced by students of several ages and proficiency levels of Swedish. The error categorization mainly focused on spelling, split compounds, merged words and simple

grammatical error. By the end of the study, Stymne *et al.* (2017) verified that “The second language learners seem to make similar mistakes as the younger children, such as confusing phonetically similar spellings (...) and writing single consonants instead of duplicate ones or the way around”. Hohn *et al.* (2016), where a comparison between the interactions of natives and non-native speakers of German in chat conversations was made, confirmed that both speakers make different kinds of mistakes and use deviations that include “orthography of German nouns and initial letters of an utterance, but also oral verb forms”. Although it was possible to see that learners had more difficulty in writing perfectly in German, Hohn *et al.* (2016) also acknowledged that “being a native speaker of a language does not necessarily correlate with high language proficiency”.

PT (source):

(15a): “**Teno** o e-mail sobre o assunto.”

EN (source):

(15b): “As soon as the feature is added, you will **recieve** an update for the app.”

Punctuation: “Punctuation is used incorrectly or is missing, or one of a pair of quotes, brackets or punctuation — e.g., “ ”, ‘ ’, (), [], { }, ¿ ?, or ¡ ! — is missing from the source text.”

PT (source):

(16a): “Obrigado[Ø] vou tentar novamente.”

EN (source):

(16b): “Please, confirm your phone number?”

Whitespace: “Whitespace is used incorrectly (there is an extra or missing whitespace).”.

PT (source):

(17a): “Gostaria de cancelar a minha assinatura[Ø]!”

EN (source):

(17b): “I am happy I was able to help you **today.Is** there anything else I can assist you with?”

Style:

Register: “The text uses **pronouns, verb forms and lexical expressions** that are not compliant with the register required for that specific text.”.

We recommend not to use this issue type in the user data because it is impossible to control the register of a user and it depends on the user and the language that is being used. However, agents have to represent the register or tone that their company decided to use in order to communicate with their clients or to represent their brand and positioning. It is crucial that the correct register is being used in the source text because we cannot expect a formal MT when something was written in an informal register in the source.

(19a): Formal English → “**Yep!** I will send you an email.”

Profanity: “A profanity is used in the source text. A profanity can be correctly written in the source language, but it can lead to problems in the translation process.”.

In some cases, the customer support experience can be stressful for both parties. This usually results in the client taking it out on the agent and ultimately using profanities to express their frustration. As opposed to a live spoken conversation, customers have to wait a certain amount of time for a response that will eventually solve their problem. In addition to having a time wait, there is also a lack of empathy in chat language that would not occur during a real life conversation (Spector, 2017). These are some of the factors that influence the use of

profanities in customer support experiences. This issue type, although named *Offensive*¹⁷, was already registered in the MQM typology and we verified some cases where a profanity was used, so for these reasons we added this issue type in the typology. Sometimes when writing profanities, it is used a method of “censorship” through the use of “x” or “*”. While at first it might seem an orthography issue, it is the case of censoring a profanity. Whether a profanity is grammatically correct or not or even censored, it can still have an impact on the target text and cause problems in the translation.

DE (source); EN (target):

(24a): Source → “Rückerstattung für ein **scheiss** game aus euren Haus...lächerlich geld zu verlangen.”

Target → “Refund for a **shy** game from your home... ridiculous money to demand.”

IT (source):

(24b): “**Caxxo**”

Wrong Language Variety: “The language variety used is not the requested one (e.g., Latin American Spanish vs. European Spanish, Brazilian Portuguese vs. European Portuguese, Traditional Chinese vs. Simplified Chinese).”

PT-PT (source):

(20a): In a PT-PT context where a PT-BR word is used → “A minha **tela** partiu-se.”

What should have been used → “O meu **ecrã** partiu-se.”

Design:

Markup: “Characters are garbled due to incorrect application of encoding.”

The encoding of some characters can be distorted in the source text. This issue type was previously named *Character Encoding*, however in both Harmonized DQF-MQM Error

¹⁷ <http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html#offensive>

Typology¹⁸ and MQM Typology¹⁹ this issue type is named *Markup*. For that reason, the name was changed.

EN (source):

(26a): “You can click on start **>** profile **>** login.”.

In this section, we will present the issue types that only concerned the TT in the MQM and Unbabel typologies and further explain why these manifested in the ST.

Accuracy:

In line with our definition of *Accuracy* and taking into account that in MQM *Addition* and *Omission* are considered within *Accuracy*, we also included these two issue types in *Accuracy*, since they affect the meaning of a sentence. It is not always straightforward to divide issue types between *Fluency* and *Accuracy* and this task was especially harder when considering monolingua data.

Addition: “An extra word or expression that is inaccurate for the sentence in a certain language.”.

Addition is an error that is usually only considered in the TT. Meanwhile, in studies on errors produced by non-native speakers, addition is a common error, especially when it comes to prepositions and articles. In Lee & Seneff (2009), it is shown that articles and prepositions present a greater difficulty to Japanese Learners of English, stating that “nouns have no determiner (“*null*”) 41% of the time, and have “*the*” and “*a*” 26% and 24% of the time, respectively”. Adding unnecessary prepositions was also common and Lee & Seneff (2009) found that the most frequently added prepositions were “*to*”, “*in*”, “*with*”, “*for*” and “*of*”. In Rozovskaya & Roth (2010), it is presented how several native speakers of multiple languages react to writing in English. Errors concerning spelling, verb form, word replacement were the most common. It is also stated that “article errors are one of the most common mistakes made

¹⁸ <https://www.taus.net/qt21-project#harmonized-error-typology>

¹⁹ <http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html#markup>

by non-native speakers of English.” (Rozovskaya & Roth, 2010), where omissions and superfluous article usage are more frequent. One example of this is when the “Superfluous *the* is usually followed by the omission of *the* and the omission of *a*.” (Rozovskaya & Roth, 2010). In the study performed by Miyata & Fujita (2021), addition was an error found in a pre-editing task of source texts, denominated as *Insertion of subject/object*²⁰.

Duplications are considered here as well because essentially a duplicated word or section in a text is an unnecessary addition to it and by joining these two issue types the IAA would increase. In the following examples we present an addition and a duplication error.

PT (source):

(1a): “Desde **de** já agradeço a atenção.”

(1b): “As **as** palavras sumiram da tela do meu relógio.”

Omission: “When a word or expression is missing and it is essential for the understanding of a sentence. An *Omission* issue type is used when a preposition, a conjunction, a determiner, a pronoun, an auxiliary verb or a word belonging to any morphological category is missing in the source text.”.

In the Unbabel Error Typology, this issue type is very fine-grained and divided into several child issue types, namely *Omitted Preposition*, *Omitted Conjunction*, *Omitted Determiner*, *Omitted Pronoun*, *Omitted Auxiliary Verb* and *Other POS Omitted*. As previously stated in *Wrong Function Word*, it was decided to go with a more general issue type in order to help the annotation process and increase the IAA, instead of having multiple specific issue types.

As for *Omission* being mainly considered an error that only occurs in the TT, when observing errors produced by non-native speakers omissions are one of the most frequent errors. As previously mentioned, in Lee & Seneff (2009), it is shown that articles and prepositions present a greater difficulty, stating that “nouns have no determiner (“*null*”) 41% of the time”. When it came to prepositions, omissions were very frequent, although it was also usual to have confusions with a different preposition. In Rozovskaya & Roth (2010), besides having many errors concerning superfluous article usage, omission was also a very

²⁰ This can be observed in *Figure 13*.

frequent error. One example of this is the former relation between addition and omission, where the “Superfluous *the* is usually followed by the omission of *the* and the omission of *a*.” (Rozovskaya & Roth, 2010). Here are a few examples of *Omission* errors:

EN (source):

(2a): “Please, give [Ø] a few minutes.”

(2b): “The order has not [Ø] shipped yet.”

Named Entity: “A Named Entity issue type is used when a Named Entity is not canonically written, having other issue types falling upon it (such as *Capitalization*).”.

While annotating with this issue type, we could see that the end result was a wrongly translated named entity and, for that reason, the issue type to be used is *Named Entity*. We believe there was no need to annotate the cause of the error, but the end result, which is a wrong named entity.

EN (source):

(3a): “I bought an **iphone 6**.”

Lexical Selection: “The word(s) selected is(are) not appropriate for its context. The word exists in the source language, but it’s used in a wrong or strange way, resulting in an uncommon combination of words or in errors in fixed expressions.”.

This issue type only applies to content words. Content words give important information for the understanding of a sentence. Content words “consist of nouns, adjectives, verbs, and so on-words whose meaning is more or less concrete and picturable.” (Smith & Witten, 1993:2-3). This issue type accounts for false friends, synonyms, hyponyms and hypernyms that could be problematic in the MT process. In Ruzaitė *et al.* (2020), where error categories were used in the Lithuanian Learners Corpus, lexical errors were also taken into account. These errors “are restricted to word choice and meaning.”, which is very similar to the *Lexical Selection* issue type. The definition applied to this error was when “the word used by the learner is orthographically and grammatically correct but is not the most natural choice

for a native speaker in terms of word meaning and/or collocability.” (Ruzaitė *et al.*, 2020). Kraichoke (2017) also makes a very clear definition of what *Word choice* errors consist of and describes perfectly why this issue type should be used in the source text:

“A usage mistake occurs when a word or a series of words in a sentence are technically grammatically correct, but not usual in standard English. While this is an uncommon error among native speakers, ESL students often translate words from their own language and select the wrong English equivalent for the meaning they wish to express.”. (Kraichoke, 2017)

EN (source):

(4a): “I need the **providence** and postal code, please.”

Wrong Paronym: “A word that is written or pronounced in a similar way to another word, despite having completely different meanings. This can apply to words with the same or different POS.”.

Due to its similarity to the *Orthography* and the *Diacritics* issue types, we will further explain this issue type in *Section 4.2.3.1.2*.

PT (source):

(5a): “Não tenho tempo para ir **a** academia.”

In this section, we will introduce the new additions proposed in the Source Typology. These issue types came from the need during the annotation effort with the Unbabel Error Typology.

Fluency

Code Switching: “When another language besides the source language is used.”

This issue type also includes the use of loanwords, which has become a common phenomenon in several languages. We verified that this phenomenon would occur in both agent and user data. Agents would use a different language in order to create a better

relationship with their client, the following example shows how they attempt it. Interestingly, this issue type was also found in a study performed by Hammarberg & Grigonytė (2014) with the same definition proposed in this typology, as shown in *Figure 14*:

(e) code switching,

Produced form	Intended word	Spell-checker correction
fashion	fashion	fusion
exciting	exciting	excitering

Table 7. Examples of code switching.

The words in Table 7 are not intended to be Swedish in the first place, but are temporary switches from Swedish into another language, in this case English.

Figure 14. Code Switching defined by Hammarberg & Grigonytė (2014)

EN (source):

(18a): “**Merçi**. I will now forward this case.”

Style

Emoticon: “The use of an emoticon can create problems in the MT process.”.

One of the reasons why emoticons have become so frequently used in chat language is because “the usage of "emoticons" has arisen from the need for expressing feelings in a very short amount of time.” (Lind, 2012:17). Given that need, emoticons have now become natural in this content. Although these are also fairly known as emojis, in most literature on this discursive structure the denomination of emoticon was frequently used, and for that reason we will name it as such. One of the conclusions in the study by Otemuyiwa (2017) was that emoticons actually have a power to eliminate ambiguity and at the same time add emotional context to what is said. This issue type was a new addition to the typology because emoticons are one of the specificities of chat language and have become more and more frequent

nowadays. While annotating, emoticons were commonly used whether it was on agent or user data. However, sometimes these structures can have an impact on the TT. The next example was taken from actual data, where we noticed that emoticons caused a MT problem.

EN (source):

(21a): Source → “Hope you’re having a great day! 😊☀️”;

Target → “Espero que tenha um bom dia! 🙄😡”.

Conversational Marker: “Conversational Markers are common in online conversations and are very specific to each language. Given their specificity, it is preferable to tag them. They are correct in the source language, but they can potentially lead to errors in the translation process.”.

Conversational markers are presented grammatically in different classes, such as “conjunctions, interjections, adverbs, and lexicalized phrases” (Schiffrin, 1987 as cited in Cabarrão *et al.*, 2018). However, given their idiomatic nature, it is very difficult to find an equivalent in different languages, thus raising issues in the field of human and machine translation (Cabarrão *et al.*, 2018). This issue type was also an addition to the typology due to the frequency of conversational markers in this content type.

PT-BR (source):

(22a): “**Hum...** Sou um pouco limitado em informática, **rsrs** mas acredito que é essa a versão.”.

Idiomatic: “An idiomatic expression specific to the source culture is used. This can cause translation problems in the target text. This also includes jargon, which can be special words or phrases that are used by a particular group of people.”.

An idiomatic expression is a fixed expression in their respective language. These expressions are sometimes recovered by memory of the speakers without having any application of grammar rules. However, this is not the case with most of them. Only some expressions have

ungrammatical structures that are already embodied in the language, resulting in becoming accepted by its speakers. The ungrammaticality of these expressions concerns only at a syntactic level. These structures are unpredictable because there are no grammar rules that explain them or for the fact that they violate some of those rules (Oliveira, 2017). For this reason, whether an idiomatic expression is correct or not, it can still have an impact on the target text. The following example is an expression grammatically correct in English; however, it is seldom translated in a unnatural expression in other languages.

EN (source):

(23a): “Hello, **Mary’s here.**”.

Abbreviation: “The use of abbreviations can lead to problems in the MT process.”.

This was a new addition to the typology due to its high frequency in chat language. As Mattiello (2013) points out, abbreviations have become more frequent since the 19th century and one of the main reasons for it is the increasing growth of technology. The internet enabled new forums to grow and create new concepts and terms. Abbreviated terms have become popular for their rising use in sms and for the informality that comes with them (Mattiello, 2013). Abbreviations were also found in the pre-editing tasks and included in the typology presented by Miyata & Fujita (2021) as *Use/disuse of abbreviation*²¹. Abbreviations are always captured by the engines, so they can have an impact on the target text by remaining the same or create an hallucination. In the following example, two abbreviations were used in the source text and only one of them was captured and translated correctly in the target text.

PT-BR (source):

(25a): Source → “Muito obrigado pela atenção. Entrarei em **ctt** com o email **q vc** forneceu.”;

Target → “Thank you very much for your attention. I will enter **ctt** with the email you provided.”.

²¹ This can be found in *Figure 13*.

Numeration: “A numeration is used incorrectly. This includes wrong bullet points.”.

This issue type was a new addition to the typology because sometimes while counting several items, whether it was the agent or the user, instead of using characters or numbers correctly in a numeration, the character used would be something more practical in the keyboard.

(27a): Using this * instead of • .

Segmentation: “Segmented text could lead to translation errors, especially if there’s a case of a split sentence.”.

A common mistype in online conversations is accidentally clicking on the “Enter” key and this usually results in the fragmentation of the message. Since the machine does not have a way to know that a sentence is being segmented, each segment is translated individually without having its full context. This issue type was also found in the typology proposed by Miyata & Fujita (2021) as *Sentence Splitting*²², confirming that this error has taken place in a different study.

IT (source); EN (target):

(28a):



²² This can be found in *Figure 13*.

These are all the issue types present in the Source Typology. Having its structure in mind, we have 4 coarse categories that have a certain number of dependent issue types. Here are the lower-level issue types:

- Parent issue types: subcategory of the coarse categories (*Addition, Omission, Named Entity, Lexical selection, Wrong paronym, Incomplete sentence, Grammar, Typography, Code Switching, Register, Wrong Language Variety, Emoticon, Conversational Marker, Idiomatic, Profanity, Abbreviation, Markup, Numeration and Segmentation*). A total of 19 issue types;
- Daughter issue types: subcategory of the *Grammar* issue type (*Wrong Function Word, Agreement, Tense/Mood/Aspect, Wrong POS and Word Order*) and of the *Typography* issue type (*Capitalization, Diacritics, Hyphenation, Orthography, Punctuation and Whitespace*). A total of 11 issue types.

For a better understanding and visualization, the issue types hierarchy is outlined in *Figure 15*:

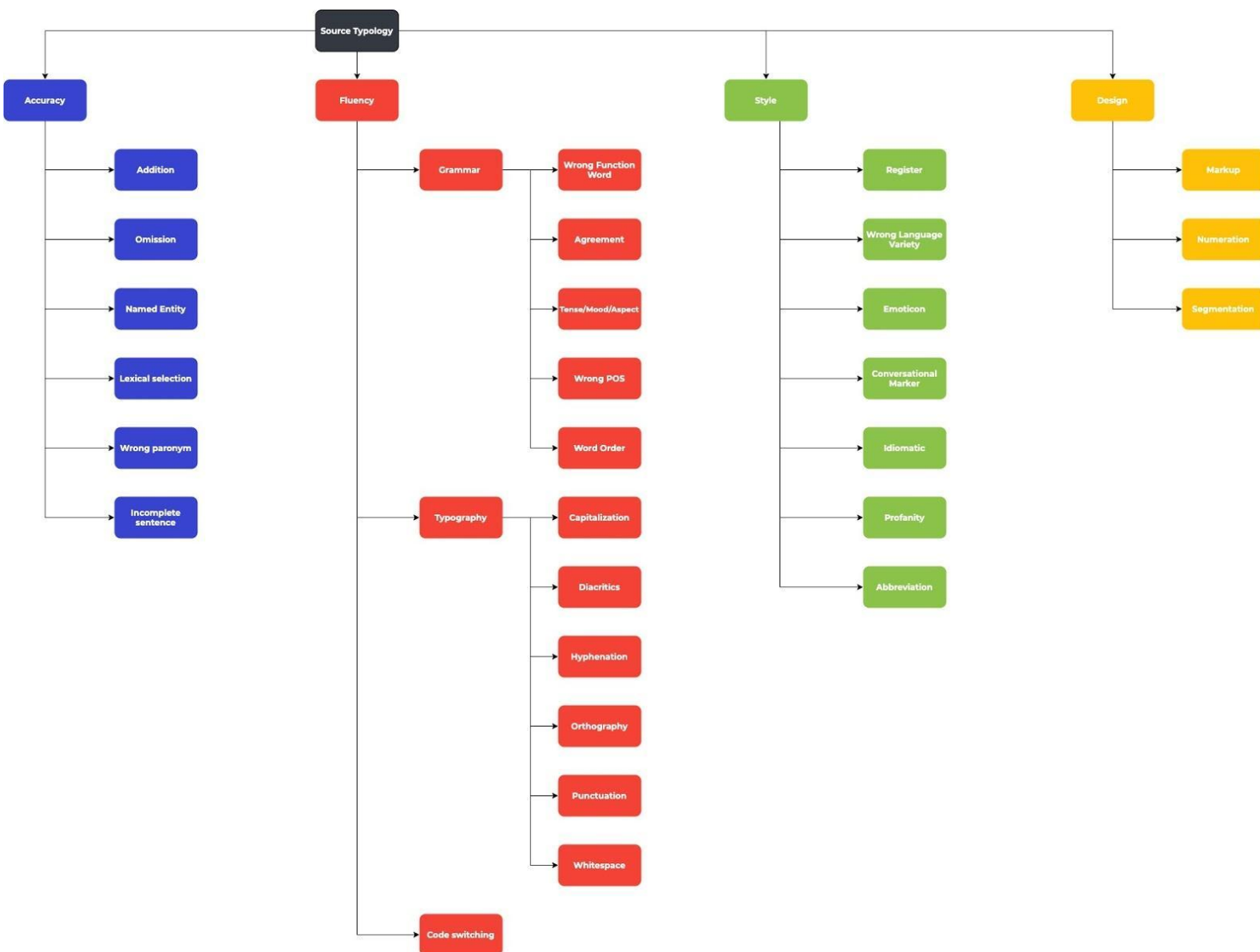


Figure 15. Source Typology diagram with 4 coarse categories, 19 parents issue types and 11 daughter issue types

4.2.3.1 Annotation guidelines

Upon having a proposed typology, we decided to write annotation guidelines. Providing these guidelines helps annotators on how to annotate the source text and how to use the Source Typology. In these guidelines, we firstly explain how to annotate with the Unbabel’s proprietary Annotation Tool by defining what is considered an error and how they

should be selected. And then we present the Source Typology with all its issue types, definitions, and a few examples for each. We have also created a section for tricky and ambiguous cases that might come up while annotating. Finally, we present two decision trees concerning doubts about the use of the issue types and the severities.

In these guidelines, we needed to clarify what is considered an error and what are the sentence structures that can cause translation problems. Besides having an error that compromises the meaning, the comprehension or any stylistic and design aspect in the source text, the source text also has linguistic structures that might have an impact on the target text. These linguistic structures might even be grammatically correct in the source text, however they can still be problematic for MT.

4.2.3.1.1 Span types

After this explanation, we also provide information on how to select an error or linguistic structure in the Annotation Tool. Having that in mind, it was important to have the definitions of a text span and a unit. A text span is the location and extension of the issue identified, while a unit is a select span. A unit can be composed of one or several words, numbers, punctuation marks, whitespaces, emoticons, and bullet points. The minimum unit that can be selected is a whole word, a whitespace, a punctuation mark, a bullet point, or an emoticon (only if it has an impact on the MT). For example, an agglutinated word that one part of it is wrong.

Formal register (EN: *It was a pleasure helping you out today*)

IT (source):

(1a) ✘ È stato un piacere aiutar[ti]_{REGISTER} oggi.

(1b) ✔ È stato un piacere [aiutarti]_{REGISTER} oggi.

In this example, the “*ti*” in “*aiutarti*” is an informal clitic. Given that the register is formal, this is an error. However, the entire span needs to be selected (“*aiutarti*”) and not just the clitic “*ti*”.

The maximum unit selected that can be selected is an entire segment that is wrong. So, if a sentence is incomplete, segmented or incomprehensible, the whole sentence needs to be selected.

IT (source):

(2a):



Here, we have an example of a segmented sentence. This would be annotated with *Segmentation*. It would not be correct to only select the first part of the sentence, the entire sentence needs to be selected.

Being aware of what a span is is essential for the annotation process because it will help the inter-annotator agreement. Although two people annotate an error with the same issue type, the span of said error can be different for each person. This is also a sign of disagreement between the annotators, so we decided to make this section as clear as possible in our guidelines. In Lommel *et al.* (2014), the precision of spans is also pointed out and perfectly exemplified. This paper held an experiment with professional translators that worked with four language pairs — EN_ES, ES_EN, EN_DE and DE_EN — to verify how different aspects affect the IAA. To demonstrate how the disagreement on an error span can affect the IAA, Lommel *et al.* (2014) show an example of two annotators that found the same issues in a sentence, however they used different spans. Here are the examples with their according span from each annotator.

ES (source):

(3a): Un **[primer año estudiante]** de PPE, que, irónicamente, había **[sido]** a Eton, dijo: “Es hija de un cerdo fascista”.

(3b): Un **[primer año estudiante de PPE]**, que, irónicamente, **[había sido]** a Eton, dijo: “Es hija de un cerdo fascista”.

The errors found in these examples were *Word Order*, *Mistranslation* and *Agreement*. In the first case (‘*primer año estudiante*’ and ‘*primer año estudiante de PPE*’), both annotators agreed that there was an issue with the word order, despite disagreeing on the extension of the issue, where one annotator felt that the part ‘*de PPE*’ also needed to change its word order, while the other annotator did not. In the case of the *Mistranslation* issue, instead of ‘*había sido*’ (‘had been’), the correct verb tense should have been “*había ido*” (‘had gone’). While one annotator used the minimal span for the issue found (3a), the other annotator used a longer one (3b). This could mean that both annotators found the same error and just used different spans or that both annotators perceived the issue in different ways resulting in highlighting the cognitively relevant span for each. (Lommel *et al.*, 2014). As for the *Agreement* error, the annotators did not agree with each other on it. While the annotator in (3a) saw it as an error, the annotator in (3b) did not. These examples found in Lommel *et al.* (2014) really show how important spans are because it might indicate a disagreement between the annotators and ambiguities in the annotation guidelines.

Once an issue is detected in the source text, it is time to highlight its corresponding text span. There are multiple types of spans, mainly continuous or discontinuous spans. Below, all spans will be explained and then exemplified.

- **Continuous spans** contain a single continuous string of text. There are two subtypes of continuous spans: single-word spans and multi-word spans.
- **Single-word spans** are whenever a word is used incorrectly and only that word should be selected.

Text	Content	Length	Error
May I have a closer look of the items?	of	1	Fluency > Grammar > Wrong Function Word

- **Multi-word spans** are when more than one word or even an expression in a continuous sequence is wrong. This would be applied to idioms or phrases that are assumed to be a single issue. So, you have to select the whole expression or words and then apply its correct issue type.

Text	Content	Length	Error
I bought the new samsung galaxy s20.	samsung galaxy s20	1	Accuracy > Named Entity

Discontinuous spans are when there is a combination of two or more separate spans that concern a single issue. There are four subtypes of discontinuous spans: delimiter spans; balanced spans, imbalanced spans and asymmetrical spans.

- **Delimiter spans** are used for typographic elements, such as punctuation, whitespaces, quotation marks, etc. This span should be used when you have, for example, unpaired quotation marks “ ”, parentheses () or even more specifically the Spanish interrogation marks ¿¿.

Text	Content	Length	Error
Click on the “Start” button.	“;”	2	Fluency > Typography > Punctuation

- **Balanced spans** are used when two disjointed but identical components are incorrect, missing or added unnecessarily.

Text	Content	Length	Error
I will really appreciated your feedback.	will; appreciated	2	Fluency > Grammar > Tense/Mood/Aspect

- **Imbalanced spans** are used to highlight two disjoint and distinct components of an issue. This type of span is appropriate for *Word Order* issues where you can highlight the misplaced items by the correct word order.

Text	Content	Length	Error
I'm sorry, we only have the black color in stock.	color;black	2	Fluency > Grammar > Word Order

- **Asymmetrical spans** are used to highlight an issue along with an element of context with which is dissonant with (the second span). These spans are used for *Agreement* issues.

Text	Content	Length	Error
The user of the app ask for instructions.	The user;ask	2	Fluency > Grammar > Agreement

Following this, we explain how the annotation process is on the platform, for example how to add and delete annotations or what steps need to be taken when the annotation process is finished.

4.2.3.1.2 Tricky cases

While annotating, some errors might be confusing or even ambiguous. For this purpose, we decided to have a section dedicated to tricky cases that might come up during the annotation process. We intended to clarify some issue types by exemplifying when they should be used.

We started with the similarity between *Omission* and *Incomplete sentence*.

The *Omission* issue type is used when a word or expression is omitted from the source text. An omission can occur at any placement of a sentence, whether it is the beginning, middle and end. When a word or expression is missing from the text, its POS is mostly evident.

EN (source):

(3a) If you want, [Ø]_{OMISSION} will send you the email. → Omitted Pronoun

(3b) If you want, [I]_{OMISSION} will send you the email.

The *Incomplete sentence* issue type is used when whole clauses are missing from a sentence. This issue type involves more than one token because it is impossible to infer the intended meaning of the original sentence. It is impossible to know how many words are missing and which is their POS. The following example is of an “if clause” whose main clause is missing.

EN (source):

(3c) I understand. If they have already done all the steps and if the console still does not read the external hard drive [Ø] [Ø] [Ø] [Ø]_{INCOMPLETE SENTENCE}. And since all the other steps were exhausted. [Ø] [Ø] [Ø] [Ø]_{INCOMPLETE SENTENCE}.

There is also some confusion with *Incomplete sentence* and *Segmentation*. In *Incomplete sentence* a significant section of a sentence is missing, while in *Segmentation* the sentence is complete but divided into more than one message. This will then result in the sentence not being translated as a whole, but as two or more different sentences.

EN (source):

(4a)



In order to not have any ambiguity, we decided to make a distinction between *Orthography*, *Wrong Paronym* and *Diacritics*. If a word is misspelled or if there is a typographical error that usually results in a non-existing word, the issue type to be used is *Orthography*.

EN (source):

(5a) May I know [qhat]ORTHOGRAPHY subscription you want to cancel?

However, when a word is misspelled, it could also result in another word that has a different meaning or in different POS. Although that word was not the one intended, the word exists in the source language. For these cases, the *Wrong Paronym* issue type should be used.

- With different meaning and same POS

EN (source):

(5b) Please be [discrete]WRONG PARONYM about what you post online.

Discrete = clearly separate or different in shape or form (**adjective**)

Discreet = careful not to cause embarrassment or attract a lot of attention (**adjective**)

- With different meaning and different POS

PT-BR (source):

(5c) O **[anuncio]**WRONG PARONYM disse que eu receberia.

Anuncio/Eu anuncio = verb “*anunciar*” in the first person singular

Anúncio = an advertisement/an announcement (**noun**)

Whenever a diacritic is used incorrectly or if a word has a missing or extra diacritic, the issue type that should be used is ***Diacritics***.

PT-BR (source):

(5d) **[Nao]**DIACRITICS, obrigada. → **[Nãõ]**DIACRITICS, obrigada.

Without its corresponding diacritic (~), the word “*Nao*” could be considered a typo. However, typos that are related to diacritics issues are not included in the ***Orthography*** issue type. Nevertheless, if the word with a missing or extra diacritic results in a grammatically correct word, then the ***Wrong Paronym*** issue type should be used.

PT-BR (source):

(5e) Estou tentando cancelar desde o dia 27 **[é]**WRONG PARONYM não consigo.

(5f) Estou tentando cancelar desde o dia 27 **[e]**WRONG PARONYM não consigo.

In the definition of ***Named Entity***, we state that when another error falls on a named entity that this issue type should be the one being used, instead of issue types such as ***Capitalization*** and so on. So, in this section we decided to exemplify the different errors that might occur in a named entity.

- The ***Capitalization*** issue type is only used when capital letters are missing or when they are being used incorrectly.

PT-BR (source):

(6a) **[olá]**_{CAPITALIZATION}, **[Bom]**_{CAPITALIZATION} dia! → **[Olá]**_{CAPITALIZATION}, **[bom]**_{CAPITALIZATION} dia!

EN (source):

(6b) Okay, **[i]**_{CAPITALIZATION} am able to help you. → Okay, **[I]**_{CAPITALIZATION} am able to help you.

If this error falls on a named entity, then the *Named Entity* issue type should be used.

EN (source):

(6c) Hello, **[jane]**_{NAMED ENTITY} ! → Hello, **[Jane]**_{NAMED ENTITY}

As explained above, the *Orthography* issue type should be used when a word is misspelled or if there is a typo.

PT-BR (source):

(6d) Estou **[cansanda]**_{ORTHOGRAPHY} de falar com pessoas que não me entendem. → Estou **[cansada]**_{ORTHOGRAPHY} de falar com pessoas que não me entendem.

If a named entity is misspelled, then use the *Named Entity* issue type.

EN (source):

(6e) The next promotion will start in **[Apiril]**_{NAMED ENTITY} and end in June. → The next promotion will start in **[April]**_{NAMED ENTITY} and end in June.

As explained previously, the *Diacritics* issue type is used when the wrong diacritic is being used or if there is a missing or extra diacritic.

PT-BR (source):

(6f) Qual é o **[horario]**_{DIACRITICS} de funcionamento do chat? → Qual é o **[horário]**_{DIACRITICS} de funcionamento do chat?

If a diacritic issue falls on a named entity, then the issue type to be used is *Named Entity*.

PT-BR (source):

(6g) Já liguei para o escritório em [Sao Paulo]_{NAMED ENTITY}. → Já liguei para o escritório em [São Paulo]_{NAMED ENTITY}.

When a punctuation mark is used incorrectly or is missing, you should use the *Punctuation* issue type.

EN (source):

(6h) No[Ø]_{PUNCTUATION}thank you[Ø]_{PUNCTUATION} → No[,]_{PUNCTUATION} thank you[.]_{PUNCTUATION}

If the spelling of a named entity involves any kind of punctuation and it is used incorrectly, the issue type to be used should be *Named Entity*.

EN (source):

(6i) Have you created an account in [Yahoo]_{NAMED ENTITY}? → Have you created an account in [Yahoo!]_{NAMED ENTITY}?

The *Hyphenation* issue type is used when a hyphen is used incorrectly or if there is a missing or extra hyphen.

EN (source):

(6j) I request you to [un-install]_{HYPHENATION} the application. → I request you to [uninstall]_{HYPHENATION} the application.

If the spelling of a named entity requires a hyphen but it is used incorrectly or missing, then the *Named Entity* issue type should be used. The same applies if the spelling of said named entity does not require a hyphen and a hyphen is added to it.

PT-BR (source):

(6k) Pode verificar toda a informação no site oficial da [MercedesBenz]_{NAMED ENTITY}. → Pode verificar toda a informação no site oficial da [Mercedes-Benz]_{NAMED ENTITY}.

The *Whitespace* issue type is used when there is an extra or missing whitespace.

PT-BR (source):

(6l) Claro[Ø]_{WHITESPACE}! Muito obrigada[Ø]_{WHITESPACE}! → Claro! Muito obrigada!

If a whitespace is used incorrectly and affects a named entity, then the *Named Entity* issue type should be used.

EN (source):

(6m) Do you have a profile in **[Linked In]**_{NAMED ENTITY}? → Do you have a profile in **[LinkedIn]**_{NAMED ENTITY}?

We decided to make a distinction between *Wrong Language Variety* and *Lexical Selection*. *Wrong Language Variety* is used when the language variety being used is not the one required in the language pair. In the following example (7a), we have the word “*esportivas*” which is exclusive to PT-BR and its equivalent in PT-PT is “*desportivas*”.

PT-PT (source):

(7a) Quero vender roupas **[esportivas]**_{WRONG LANGUAGE VARIETY}.

Lexical Selection is used when the word is not exactly appropriate for the context and another word would fit better. This is very frequent with collocations.

EN (source):

(7b) I **[lost]**_{LEXICAL SELECTION} the bus. → I **[missed]**_{LEXICAL SELECTION} the bus.

Wrong Language Variety can also be mistaken for *Code Switching*, so we decided to clarify them. The *Code Switching* issue type is used when another language, besides the source language, is used. This does not include varieties of the same language.

EN (source):

(7c) My apologies, **[Monsieur]**_{CODE SWITCHING} Alexandre.

When a language variety is wrongly used because it is not the required one, then you should use the *Wrong Language Variety* issue type.

Finally, we ended this section by distinguishing an abbreviation from an acronym and when the *Abbreviation* issue type should be used. An abbreviated word is a shortened form of a word. This is usually a result of typing too fast. When this is the case, you should use the *Abbreviation* issue type.

EN (source):

(8a) Please, provide your **[acct]**_{ABBREVIATION}. → account

An acronym is often mistaken for an abbreviation. An acronym is an abbreviation consisting of the first letters of each word in the name of something, pronounced as a word.

EN (source):

(8b) **NATO** was founded in 1949, right after the II World War.

NATO = North Atlantic Treaty Organization

If an acronym is written incorrectly, then this falls on the *Named Entity* issue type.

EN (source):

(8c) **[ANTO]**_{NAMED ENTITY} was founded in 1949, right after the II World War.

These are the tricky cases explained and exemplified in the annotation guidelines. We tried to distinguish issue types that seemed similar. If any other doubts about two issue types come up during the annotation process, this section will be expanded accordingly.

4.2.3.1.3 Severities

In the annotation guidelines, we have also provided further information about the severities that are used with the Source Typology. A severity level indicates how serious an error is. Having different levels of severity helps to predict the impact of the error in the source text and also to calculate MQM. In the Source Typology, there are going to be four different severity levels:

- Critical (10 points)
- Major (5 points)
- Minor (1 point)
- Neutral (0 points)

These severity levels will be used according to the way they affect the *Accuracy*, *Fluency*, *Style* and *Design* of the source text. The higher the severity level, the more the quality of the text is going to be affected. In the annotation guidelines, we explain each severity and when it should be applied and finally give examples. Here are the severities:

Critical: an error should be cataloged as critical when it is:

- An information that may carry health, safety, legal or financial implications;
- A violation of geopolitical usage guidelines;
- Misrepresentation of the concerned company and their respective product/service;
- Content completely inappropriate to its target audience.

Some examples of critical errors include:

- A word selection that affects the meaning of the text or that has a negative influence on the reader towards a certain product or service.

IT (source):

(9a) [Tifare]_{LEXICAL SELECTION} l'ordine. → [Rifare]_{LEXICAL SELECTION} l'ordine.

In this example, “*tifare*” means “to cheer” in the context of football and “*rifare*” means “do it again”. So, the meaning of the sentence changes completely.

- An incomplete sentence that affects the message of the text.

EN (source):

(9b) Once you’ve completed this step, you [Ø] [Ø] [Ø] [Ø]INCOMPLETE SENTENCE.

Major: an error should be catalogued as major when there is:

- Misleading information;
- Change of meaning;
- Register wrongly used.

Some examples of major errors include:

- Agreement
- Named Entity (if its meaning is compromised)
- Tense/Mood/Aspect
- Omission
- Word Order
- Wrong Function Word
- Register
- Lexical Selection (only when the word it is not appropriate to its context)
- Wrong Paronym
- Wrong POS
- Markup

PT-PT (source) formal register:

(10a) Para ver mais informações, [vai]REGISTER ao nosso site. → Para ver mais informações, [vá]REGISTER ao nosso site.

EN (source):

(10b) If you want, [Ø]OMISSION will send you all the information to your second email.

→ If you want, [I]OMISSION will send you all the information to your second email.

Minor: an error should be catalogued as minor when there are:

- Minor aspects that can be solved with proofreading.

EN (source):

(11a) Perfect [Ø]WHITESPACE! → Perfect!

(11b) I cannot make a commitment regarding this as [thye]ORTHOGRAPHY development team has not yet provided any update about this issue. → I cannot make a commitment regarding this as [the]ORTHOGRAPHY development team has not yet provided any update about this issue.

Neutral: with the new issue types added, we could see that there were linguistic structures that were correct in the source text, but could lead to problems in the target text. For that reason, this severity level is not used for errors but for linguistic structures that might have an impact on the MT process. The neutral severity falls exclusively upon these issue types:

- Emoticon
- Code Switching
- Segmentation
- Conversational Marker
- Idiomatic
- Profanity
- Lexical Selection (only when tagging a false friend)
- Wrong Language Variety
- Abbreviation

PT-BR (source):

(12a) Dá [pra]ABBREVIATION ver que não está a 100%.

EN (source):

(12b) Thanks! [:)]_{EMOTICON}

4.2.3.1.4 Decision trees

Finally, the annotation guidelines are concluded with the decision trees that we provided. Decision tree is a support tool that uses a tree-like model of several decisions and their resulting consequences. These decision trees were made to help annotators as they are learning tools that aid to clear any doubts or concerns that might arise in the annotation process, especially if there is an issue where the answer is not immediately clear (Burchardt & Lommel, 2014). In order to better assist the annotators, it was decided to construct two decision trees — one concerning the issue types of the Source Typology and the other concerning the severities used in the said typology.

As Burchardt & Lommel (2014) stated, decision trees should be presented in a hierarchical mode, going from a more general to a more specific type. In the decision tree concerning the issue types, we started with the coarse categories, then progressed into its parent and child issue types. Both trees were built with an elimination process in-mind, where one question concerning an issue is asked and if the answer is ‘no’, then the annotator would go to the next question concerning another issue type. Not only are the issue types belonging to the same category connected, but they are also connected with the other categories if that is the case. We decided to follow the same structure displayed in the tricky cases sections, specifically the issue types that could be considered more ambiguous.

Firstly, we presented a decision tree concerning the usage of the issue types present in the Source Typology.

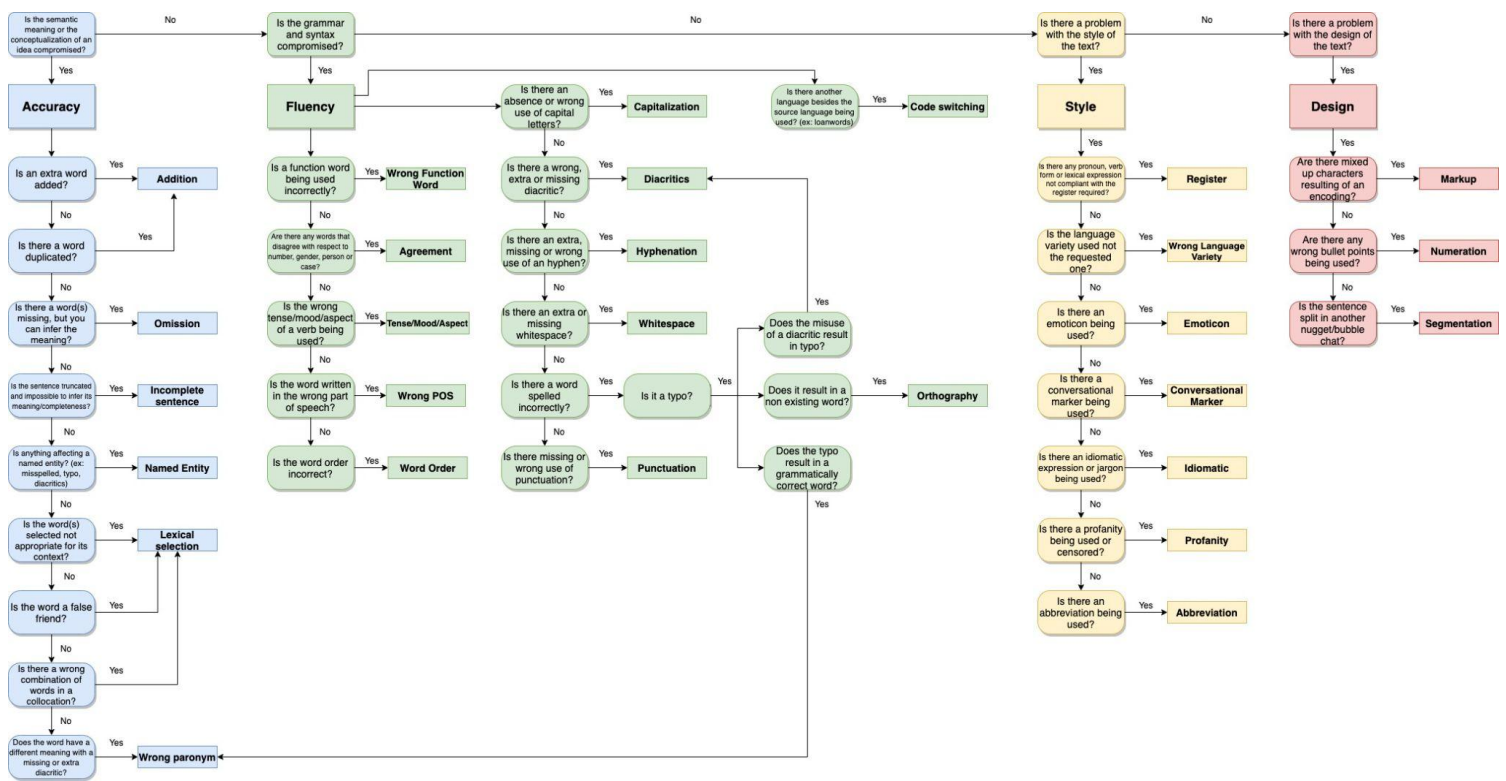


Figure 16. Source Typology decision tree

Additionally, we presented a decision tree concerning the severities in case the annotators have any doubts of which severity is more suitable with the error they are annotating, highlighting its differences.

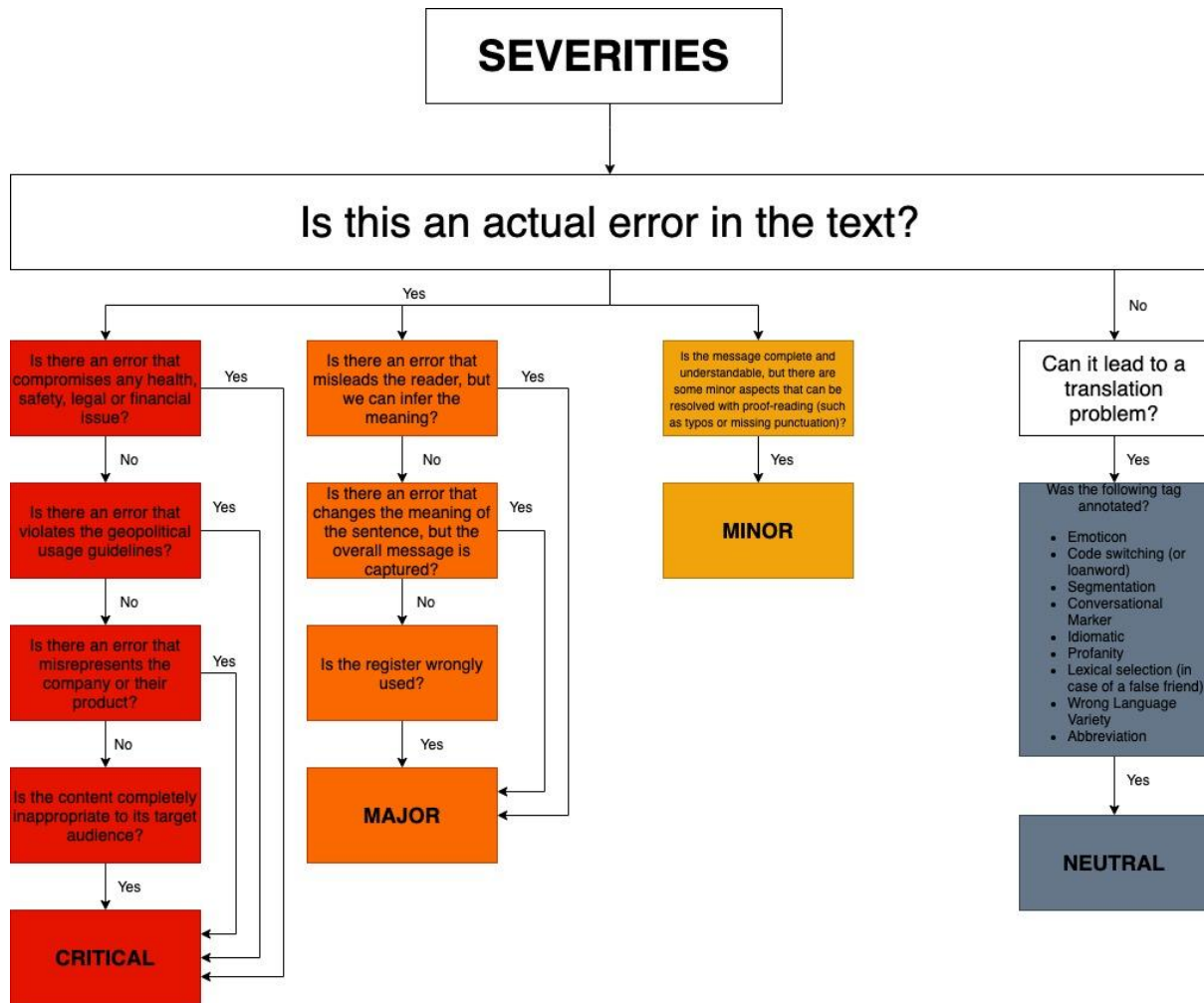


Figure 17. Severities' decision tree

5. Results and discussion

With a proposed typology and its annotation guidelines, it was finally time to test it with more languages and with different data. This would allow us to see if there was anything that could be improved in the Source Typology, for example better explanations of the issue types and if there is any improvement concerning the new issue types and the *Neutral* severity.

With these annotation guidelines, we now present an internal pilot with the Source Typology, which was divided into three case studies: PT-BR_EN inbounds, Agent annotation and Multilingual internal pilot.

5.1 PT-BR_EN inbounds

In the first case study, PT-BR_EN inbounds, the source language was Brazilian Portuguese and the target language was English. The data to be annotated was randomly selected, resulting in different clients. After finishing the annotation, the MQM was calculated. Not only the MQM of the ST, but also of the TT with multiple MT systems available online. The annotations of the TT were performed with the Unbabel Error Typology for consistency reasons. Once the MQM score was very low with the MT System 1, it was decided to also annotate the TT with the MT System 2 and MT System 3 to compare the final results. The focus of this study was to see the impact that the ST actually has on the TT.

	Number of words	Neutral	Minor	Major	Critical	MQM
Source	2909	27	497	62	0	72.26
MT System 1	2874	N/A	52	175	85	29.05
MT System 2	3003	N/A	86	50	36	38.96
MT SYSTEM 3	2998	N/A	66	98	49	51.37

Table 3. PT-BR_EN inbounds' annotation results

As can it is shown in *Table 3*, the source MQM score, although low, is much higher than any target MQM score. One of the reasons for this was because the number of *Major* and *Critical* errors is more recurrent in the target text, while in the source text there were not any *Critical* errors. To demonstrate the differences between the engines, it will be presented below an example of a sentence translated by the three engines.

Source	OLA BOA NOITE ESTOU INICIANDO NO RAMO DE DROP E GOSTARIA DE SABER SE ASSIM QUE E ESCOLHER OS PRODUTOS EU POSSO JOGAR DIRETO NA LOJA ,OU TENHO QUE ENTRAR EM CONTATO COM O FORNECEDOR DO [ORG] ²³ PRIMEIRO, ESTOU TOTALMENTE PERDIDA
MT System 1	Ola BOA NOITE was starting in the DROP RAMO AND GOSTARIA OF KNOWING AS AS SIGNED AND SHOULD THE PRODUCTS I TAKE TO [ORG] ²⁴ AliExpress TAKE
MT System 2	OLA GOOD NIGHT I AM STARTING IN THE DROP BRANCH AND WOULD LIKE TO KNOW IF SO AND CHOOSING THE PRODUCTS I CAN PLAY DIRECTLY IN THE STORE, OR I HAVE TO CONTACT THE FIRST [ORG] ²⁵ SUPPLIER, I AM TOTALLY LOST
MT System 3	HELLO GOOD EVENING I'M STARTING IN THE DROP BUSINESS AND I WOULD LIKE TO KNOW IF SO AND CHOOSE THE PRODUCTS I CAN PLAY DIRECTLY IN THE STORE, OR I HAVE TO CONTACT THE SUPPLIER OF [ORG] ²⁶ FIRST, I AM TOTALLY LOST

Table 4. PT-BR_EN inbounds' target examples

Here, we have an example of a sentence that was fully capitalized in the source text and in each engine different errors occurred. With the MT System 1, some words were left untranslated, resulting in a mixture of Brazilian Portuguese and English (“*AND GOSTARIA OF KNOWING*”). The same happened with MT System 2, although it just left one word untranslated (“*OLA*”). Another occurrence in the MT System 1 was cutting the sentence by half, without acknowledging the rest of it, which did not happen with the other two engines. The MT System 3 did not have as many errors as the other two engines, but it still had errors that were related to the source text, such as missing punctuation and diacritics.

To understand better the relation between the source and target, an alignment between the annotation results was made. This alignment was only done with the target translated by the MT System 1.

²³ Anonymized organization

²⁴ Anonymized organization

²⁵ Anonymized organization

²⁶ Anonymized organization

Same errors found in both source and target	Source errors that originated different target errors	Neutral issue types on the source that had an impact in the target
34	29	9

Table 5. PT-BR_EN source and target alignment

While aligning the source with the target, it was possible to verify that some errors found in the source would be transferred in the target and in other cases the source errors could originate different errors in the same sentence. It was also decided to check the effectiveness of the *Neutral* severity and if it did have an impact in the MT process or not. From 27 issue types, 9 of them were problematic in the target text. In order to facilitate the understanding of the alignment, several examples will be provided. Firstly, an example of the same errors found in both the source and target:

Source	Source error	Source typology error	Source severity	Target	Target error	Target typology error	Target severity
a garota do trem	a garota do trem/	Named Entity/Punctuation	Minor/Minor	The train girl	The train girl	Named Entity	Critical
A FILHA DO CONDE	A FILHA DO CONDE	Named Entity/Segmentation	Minor/Major	The FILE OF THE CONDE	The FILE OF THE CONDE	Named Entity	Critical
A DAMA MAIS DESEJADA	A DAMA MAIS DESEJADA /A DAMA MAIS DESEJADA	Named Entity/Segmentation	Minor/Major	The MOST DESIGNED DAMAGE	The MOST DESIGNED DAMAGE	Named Entity	Critical

Table 6. PT-BR_inbounds' same errors found both in the source and target

As it can be seen in *Table 6*, when *Typography* errors fall on a named entity, despite not being very problematic in the source, they can interfere with the target text and originate errors of a higher severity level. In these examples, we can see that the use or misuse of capitalization in a named entity will compromise the translation of said named entity in the target language, by translating it literally or by hallucinating its name completely. In *Table 7*, there will be source errors that originate different target errors:

Source	Source error	Source typology error	Source severity	Target	Target error	Target typology error	Target severity
COMPREI UM LIVRO ONTEM PARA MINHA FILHA E NA HORA QUE COLOCO PARA BAIXAR ELE DA ERRO .	COMPREI UM LIVRO ONTEM PARA MINHA FILHA E NA HORA QUE COLOCO PARA BAIXAR ELE DA ERRO/DA	Capitaliz ation/Wr ong Paronym /Whitesp ace	Minor/ Minor/ Minor	I bought A FREE SHIPPING FOR MY FILE AND IN THE TIME I COLOKED TO BOX IT FROM THE ERROR.	FILE/C OLOKE D/A FREE SHIPPI NG/TO/ FROM	Lexical Selection/ Unintellig ible/Addit ion/Wrong Prepositio n/Omitted Determine r/Lexical Selection	Critical/C ritical/Cri tical/Maj or/Major/ Critical

Table 7. PT-BR_EN inbounds' source errors that originated different target errors

In this example, we can see that a simple sentence with a *Capitalization* and *Wrong Paronym* error originated several other errors with a much higher severity. The use of capitalization had an impact on the target text by translating incorrectly the noun “*FILHA*” to “*FILE*” (when it should be “daughter”) or by creating a new word (“*COLOKED*”) in the target language. By having a wrong paronym in the source text, where “*DA*” (preposition) should

have a diacritic/accent in order to distinguish it from “*dá*” (verb), this was translated as preposition when it should have been a verb. Finally, in *Table 8*, we will present neutral issue types that were used in the source text that had impact in the target text:

Source	Source error	Source typology error	Source severity	Target	Target error	Target typology error	Target severity
Abs	Abs/	Abbreviation/Punctuation	Neutral/Minor	Abs	Abs	Unintelligible	Critical
Responda ao que eu pergunto sff	/sff	Punctuation/Abbreviation	Minor(Neutral)	Answer what I ask sff	/sff	Omitted Preposition/Unintelligible	Major/Critical

Table 8. PT-BR_IN inbounds' neutral issue types on the source that had an impact on the target

In *Table 8*, there are two examples of linguistic structures with a *Neutral* severity that were problematic on the target text. In these cases specifically, there was the use of abbreviations in the source text which were not captured by the engine, thus resulting in not being translated in the target language where these abbreviations have no meaning. So, this is a perfect example of the impact that the *Neutral* severity might have on the target text, where it could result in a *Critical* error.

In *Figures 18* and *19*, there will be the breakdown of all the errors and severities used in both source and target annotations:

Sample PT-BR_EN (inbound): source errors and severities breakdown

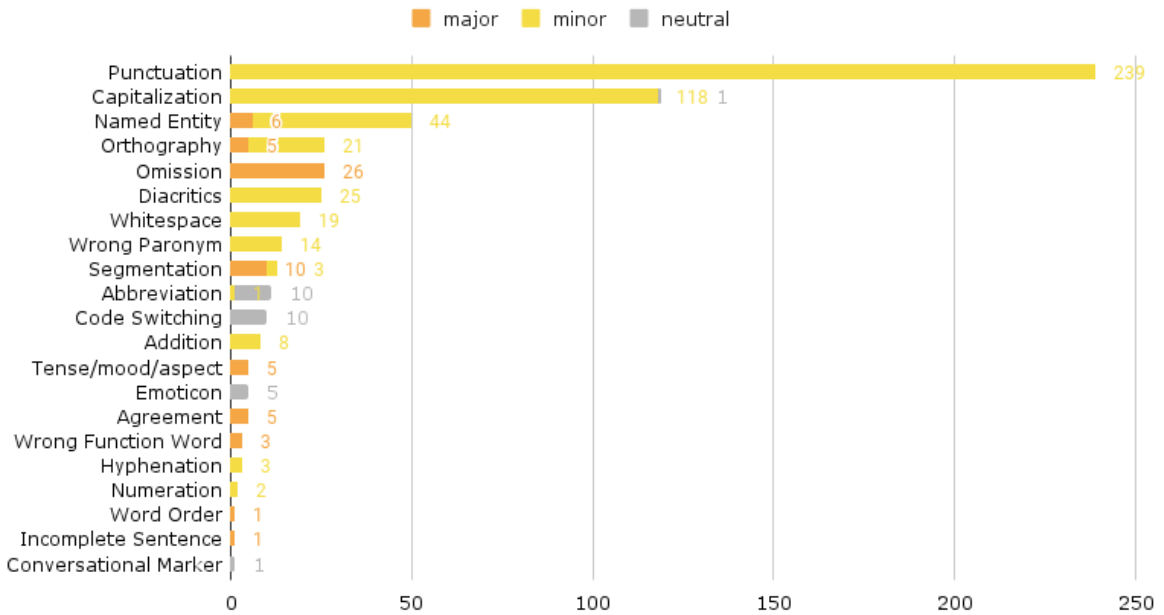


Figure 18. PT-BR_EN (inbound): source errors and severities breakdown

Sample PT-BR_EN (inbound): target errors and severities breakdown

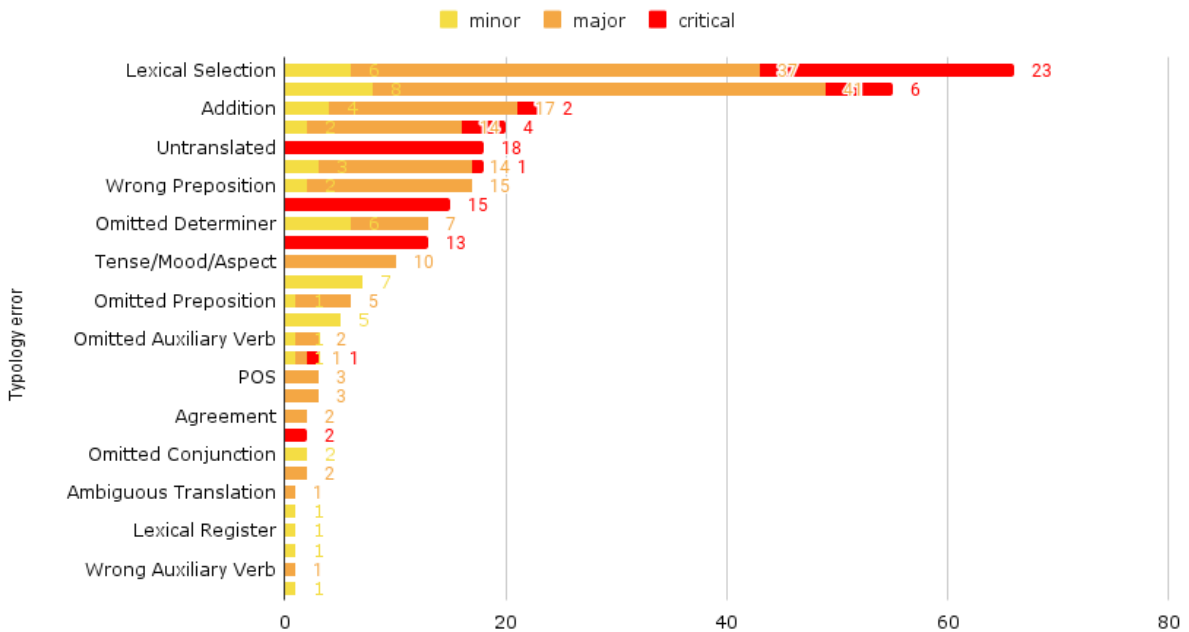


Figure 19. PT-BR_EN (inbound): target errors and severities breakdown

While in the source annotation, the most common errors were *Typography* errors such as *Punctuation*, *Capitalization*, and *Orthography*, which also fell on the *Named Entity* issue type, in the target annotation, the most common errors were *Lexical*, *Other POS Omitted*, *Addition*, *Overly Literal* and *Word Order*. With these figures it is easier to see the difference between the severities used in each one, where the *Minor* severity was predominantly used in the source annotation and the *Major* and *Critical* severities were mainly used in the target annotation.

5.2 Agent Annotation

In the second case study, Agent Annotation, the source language was English and the target language was French. In this case study, it was decided to only annotate data from a single client so that we could verify the consistency of the translations. The same procedure of the PT-BR_EN inbounds was performed with the MQM evaluation, although in this case study the target annotation was with the Unbabel MT due to a high score.

	Number of words	Neutral	Minor	Major	Critical	MQM
Source	9848	17	409	341	0	78.53
Target	10,707	0	211	226	2	87.41

Table 9. Agent Annotation's results

As it is presented in *Table 9*, the target MQM score is slightly higher than the source MQM score. Despite annotating two *Critical* errors in the target text, the number of *Minor* and *Major* errors is lower than the ones annotated in the source text. In *Figures 20* and *21*, there will be a breakdown of all the errors and severities used in both source and target annotations:

Agent annotation: source errors and severities breakdown

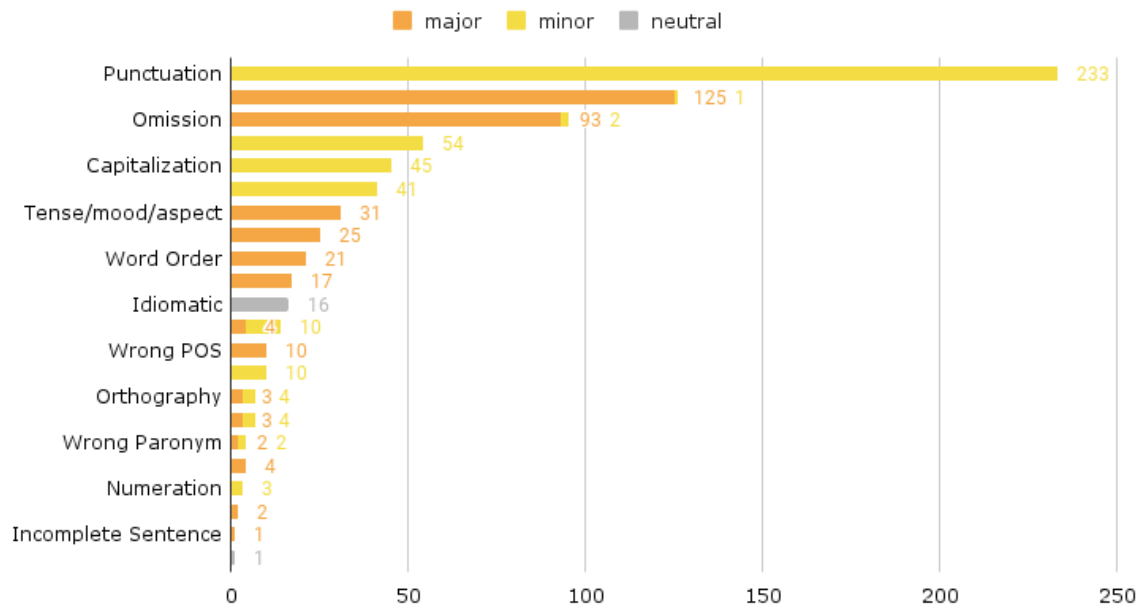


Figure 20. Agent annotation: source errors and severities breakdown

Agent annotation: target errors and severities breakdown

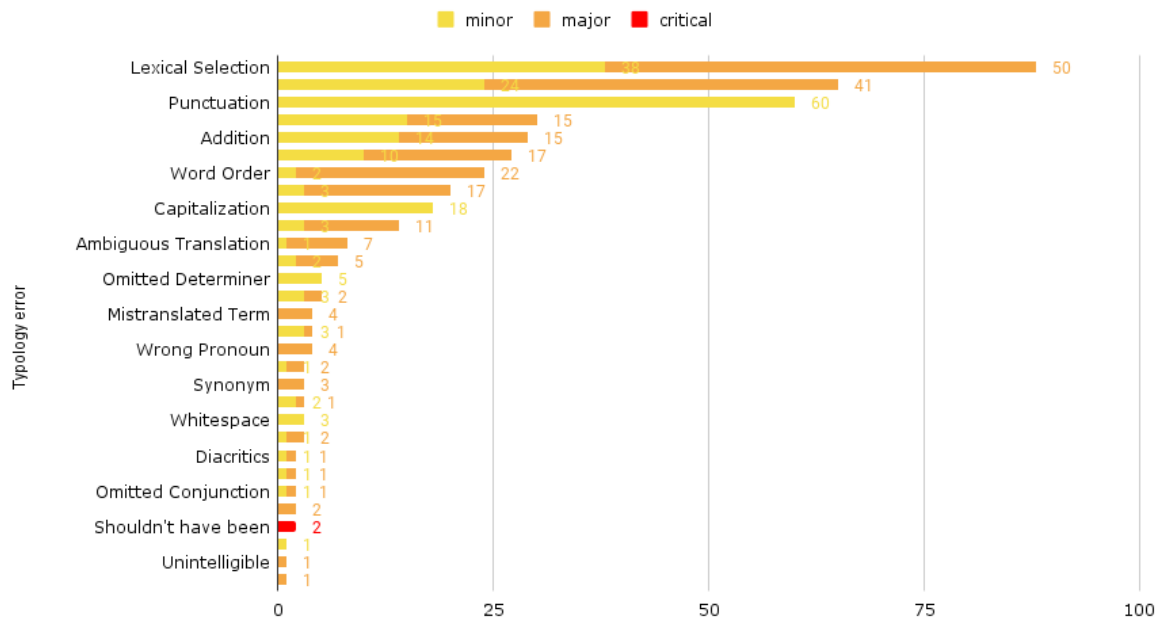


Figure 21. Agent annotation: target errors and severities breakdown

The most common errors found in the source were *Punctuation*, *Register*, *Omission*, *Whitespace* and *Addition*, while the most common errors found in target were *Lexical Selection*, *Overly Literal*, *Punctuation*, *Tense/Mood/Aspect*, and *Addition*. In both the source and target, it was used mainly the *Minor* and *Major* severities. However, in the source the *Neutral* severity was used for the *Idiomatic* and *Conversation Marker* issue types. It was also decided to make an alignment between the source and target to check if the source errors had any impact on the target text.

Same errors found in both source and target	Source errors that originated different target errors	Neutral issue types that an impact on the target
59	40	0

Table 10. Agent annotation's source and target alignment

By aligning the source with the target, it was possible to see that the source had a lot of impact on the target text with 59 errors that were found in both and 40 source errors that originated in other errors in the target. However, in this case study the *Neutral* severity had no impact on the target. So, for the other two cases there will be presented examples in order to better understand the reason why they happened. First an example of the same errors found in both source and target:

Source	Source error	Source typology error	Source severity	Target	Target error	Target typology error	Target severity
I check the details for the other order number DEVICEID-0, it shows here the items are already delivered last 12/01/2020 and you did not receive this two items as well, correct?	check/are / /this	Tense/Mood/Aspect/Tense/Mood/Aspect/Omission/Agreement	Major/Major/Major/Major	Je vérifie les détails de l'autre numéro de commande DEVICEID-0, cela montre ici que les articles sont déjà livrés le dernier 12/01/2020 et que vous n'avez pas également reçu ces deux articles, n'est-ce pas ?	sont/12/01/2020/dernier/égalelement	Tense/Mood/Aspect/Word Order/Tense/Mood/Aspect	Major/Minor/Major

Table 11. Agent annotation's same errors found in both source and target

In this example, we have the issue type *Tense/Mood/Aspect* in the source (“are”) that was also found in the target text (“sont”). The same severity, *Major*, was used in both source and target. In *Table 12*, it will be presented source errors that originated different target errors:

Source	Source error	Source typology error	Source severity	Target	Target error	Target typology error	Target severity
I understand you want to receive this shoe and will be delivered it to you, Allo me to check it within 2-3 minutes?	delivered/, /Allo	Orthography/ Punctuation/ Orthography	Minor/ Minor/ Minor	Je comprends que vous souhaitez recevoir cette chaussure et qu'elle vous sera livrée, Allo moi pour la vérifier dans les 2-3 minutes ?	livrée/Allo	Lexical Selection/ Untranslated	Critical/ Critical

Table 12. Agent annotation's source errors that originated different target errors

In this example, we have two *Orthography* errors that were annotated with the *Minor* severity in the source text, while in the target text these errors formed in *Lexical Selection* and *Untranslated* errors with a *Critical* severity.

5.3 Multilingual internal pilot

We decided to run a multilingual internal pilot to study the efficiency of the Source Typology. As the name suggests, this case study was performed internally at Unbabel and it was only possible thanks to its workers that volunteered to test it. The source languages annotated in this experiment were Dutch, Polish, Romanian, Brazilian Portuguese, Italian,

Spanish, German and English. There were at least one or two volunteers in each language. This pilot was divided into user annotations and agent annotations. The user annotations were in all languages, except for English, that was agent annotations. The results of the annotations can be found in *Tables 13* and *14*:

USER ANNOTATION RESULTS						
Source language	Number of words	Neutral	Minor	Major	Critical	MQM
Dutch (NL)	2884	27	103	22	5	90.88
Italian (IT)	977	24	67	3	4	87.51
Spanish (ES)	1560	13	153	14	0	85.71
Brazilian Portuguese (PT_BR)	1838	35	185	31	0	81.5
Romanian (RO)	536	2	106	14	2	63.43
Polish (PL)	1519	9	125	91	0	61.82
German (DE)	1942	33	186	146	0	53.09

Table 13. Internal pilot User Annotation results

As can be seen in *Table 13*, the user MQM scores vary from language to language, with Dutch having the highest MQM (90.88) and German with the lowest MQM score (53.09).

AGENT ANNOTATION RESULTS						
Source language	Number of words	Neutral	Minor	Major	Critical	MQM
English (EN)	2842	52	193	43	0	85.46

Table 14. Internal pilot Agent Annotation results

In *Table 14*, we have the annotation results from the agent data that only consisted of the English language, whose source MQM (85.46) was very high.

5.3.1 Neutral structures analysis

We were able to use the main novelty of the Source Typology, which is the *Neutral* severity, and for that reason the focus is going to be mainly on it and it will then be verified if it had any impact on the target text. First, it was decided to check how many segments were annotated and compare it with the total number of annotated errors and then check how many of them were annotated as *Neutral*.

Source language	Source segments	Total annotated errors	Neutral structures
DE	236	365	33
EN	294	287	52
PT_BR	212	252	35
PL	140	225	9
ES	127	180	13
NL	357	157	27
RO	61	124	2
IT	159	87	24

Table 15. Total number of annotated errors and neutral structures found in the source

In almost every language, except for Italian and Dutch, the number of errors is much higher than the number of segments, which would mean having multiple errors in a sentence. The number of neutral structures is just a small fraction of the total number of errors, however these linguistic structures might still have an impact on the target language.

Before analyzing the impact of Neutral structures on the MT, we decided to look into the issue types that were used with the *Neutral* severity. On an important note, it was observed that every annotated language has linguistic structures that are correct but were annotated as *Neutral*, as they might have some influence in the translation, which will be displayed in this section.

Neutral typology structures

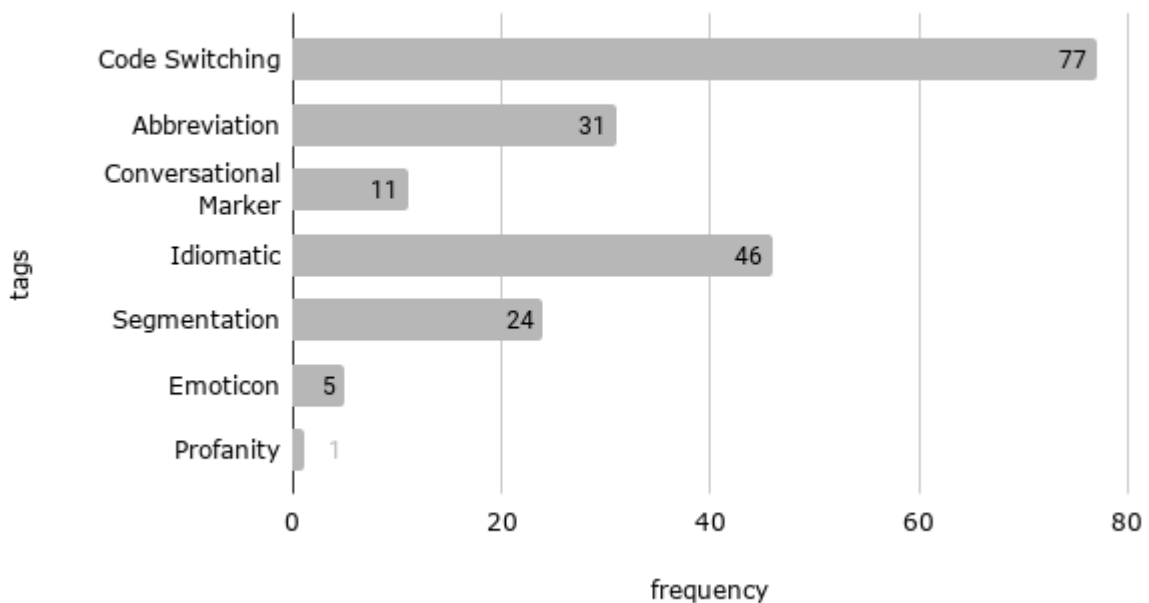


Figure 22. Internal pilot's annotated Neutral typology structures

The issue types used in all languages were, by higher to lower frequency, *Code Switching*, *Idiomatic*, *Abbreviation*, *Segmentation*, *Conversational Marker*, *Emoticon*, and *Profanity*.

Neutral structures in *User's* data

In the *User's* data, the linguistic structures annotated as neutral were *Code Switching* (Table 16), *Idiomatic* (Tables 17 and 18), *Abbreviation* (Tables 19 and 20), *Segmentation*, *Conversational Marker* (Tables 21 and 22), *Emoticon* (Tables 23 and 24) and *Profanity* (Table 25). *Code Switching* is annotated whenever another language besides the source language is being used. In the case of *Code Switching*, the examples below show their impact on the machine translation into the target language. In such cases, the language being used (besides the source language) was mostly English. Since the target language was also English, the MT left it the same. However, two segments were incorrectly written in English in the source, thus remaining the same in the target text and incorrect in the target language. These examples (1a and 1b) are illustrated in Table 16. The source texts presented may have other errors, however we will only highlight the neutral structures that had an impact on the target text.

EN (source):

(1a): [he platforms]_{Code Switching}

PL (source):

(1b): Proszę o przejście do zakładki [resolution ceneter]_{Code Switching}, bądź też sprawdzić skrzynkę e-maila, oraz zapoznać się z treścią wiadomości.

Source language	Source text	Linguistic structure	MT Target text	Post-edited Target text
DE	he platforms	he platforms	he platforms	The platforms.
PL	Proszę o przejście do zakładki resolution ceneter , bądź też sprawdzić skrzynkę e-maila, oraz zapoznać się z treścią wiadomości.	resolution ceneter	Please go to the resolution ceneter tab, also check the scan of the e-mail and see the message content.	Please go to the resolution center tab, also check the scan of the e-mail and the message content.

Table 16. Code Switching examples

We had annotations with *Idiomatic* that had an impact on the MT, which will be illustrated with some examples below. *Idiomatic* is annotated when an idiomatic expression specific to the source culture is being used.

IT (source):

(2a): già provato...[mi sa]_{Idiomatic} che sono costretto ad una riparazione a questo punto

PT-BR (source):

(2b): [Super entendo]_{Idiomatic} essa variação de valores do real vs dólar, mas a minha dúvida é pq o [ORGANIZATION] aparece com um

IT (source):

(2c): ciao ho un [PRODUCT] e mi dice, dopo averlo acceso [dopo secoli]_{Idiomatic}, che c e un aggiornamento.

In the example (2a) there is an Italian idiomatic expression, “*mi sa*”, that was translated incorrectly. It should have been translated to “*I think*” or “*I guess*”, instead of “*I know*”.

Source language	Source text	Linguistic structure	MT Target text	Post-edited Target text
IT	già provato... mi sa che sono costretto ad una riparazione a questo punto	mi sa	Already tried... I know I'm forced to repair at this point	Already tried... I guess I'm forced to make a repair at this point.

Table 17. Italian idiomatic expression translated incorrectly

The examples (2b) and (2c) had idiomatic expressions that were translated to the target language but still sounded unnatural in that language (“*super understand*” and “*after centuries*”).

Source language	Source text	Linguistic structure	MT Target text	Post-edited Target text
PT_BR	Super entendo essa variação de valores do real vs dólar, mas a minha duvida é pq o [ORGANIZATION] aparece com um	Super entendo	Super understand this variation of values of the real vs dollar, but my doubt is pq the [ORGANIZATION] appears with a	I totally understand this variation of values of the real vs dollar, but my question is why does [ORGANIZATION] appear with a divergent value when it is placed in the application? Find this variation attached!

IT	ciao ho un [PRODUCT] e mi dice, dopo averlo acceso dopo secoli , che c e un aggiornamento.	dopo secoli	Hello I have a [PRODUCT] and it tells me, after having turned it on after centuries , that there is an update.	Hello, I have a [PRODUCT] that I didn't turn on in ages and now it gives a message saying that there's an update available.
-----------	--	-------------	---	--

Table 18. Unnatural idiomatic expressions

Abbreviations are very common in chat language and very specific to each language, and users tend to use them more frequently than ever before. This could be problematic in the target text because the MT behavior is different when it comes to abbreviations. In a total of 31 abbreviations, only 12 of them were translated correctly in the target language. Here are two examples of abbreviations that had an impact on the target text.

PT-BR (source):

(3a): Olá, tenho uma compra que está pendente [pq]_{Abbreviation} foi rejeitada no momento que usei o cartão.

NL (source):

(3b): En dat boek van kolletje heb ik destijds [oa]_{Abbreviation} gekocht, maar die kan ik dus niet openen

Most of the abbreviations used in the source text were left untranslated in the target text, hence being incomprehensible in English (as the example (3a)). In this particular case, we have a Portuguese abbreviation for the word “*porque*” (which means “because” in English), that remained the same in the target text.

Source language	Source text	Linguistic structure	MT Target text	Post-edited Target text
PT_BR	Olá, tenho uma compra que está pendente pq foi rejeitada no momento que usei o cartão.	pq	Hello, I have a purchase that is pending pq was rejected at the time I used the card.	Hello, I have a purchase that is pending because it was rejected the moment I used the card.

Table 19. Portuguese abbreviation example

In the example (3b) we have a Dutch abbreviation that was omitted in the translation.

Source language	Source text	Linguistic structure	MT Target text	Post-edited Target text
NL	En dat boek van kolletje heb ik destijds oa gekocht, maar die kan ik dus niet openen	oa	And I bought that book of collet at the time [] , but I can't open it	And at the time I bought that book by Kolletje, among other things , but I couldn't open it.

Table 20. Dutch abbreviation example

The segmentation of a text or paragraph can have an impact on the target text. To cover this, we have *Segmentation*, that is used whenever we have a segmented text. Writing in chat, ultimately results in the fragmentation of ideas. This is mostly due to the time restrictions of chat language and how this “affects chat language and despite it being a written medium it is highly fragmented” (Lind, 2012:17). In the case of users' data, they are less affected by time itself than agents are, however they are still affected by it. This could result in the fragmentation of a sentence or an idea into several chat messages. If this is the case in the source text, then there is a chance that the target text might be affected. The machine does not translate the entire chat conversation as a whole, but it rather translates segment by segment,

which leads to a great loss of context and poor syntax and grammar. For these cases, we introduced the *Segmentation* issue type. In order to present some instances of this, we will display some examples below. In the first example presented below is a sentence written in Italian that was segmented into three different chat messages. When looking at the sentences standalone, their translations look good but, when in context, we can see that there are errors related to the adverb *solo* translated into *only* and the capitalization of *Account*. We also have a minor *Word Order* error with the adverb “only”. Its translation was “I see only”, however if we look at it with its context in mind, its correct translation is “I only see”.

IT (source); EN (target):

Source	MT Target	Post-edited Target text
vedo solo	I see only	
account	Account	I only see account or “sign in”.
O “accedi”	Or “access”	

Another example was a fragmented greeting formula which resulted in a poor translation of the second chat message “*I now turn to the situation.*”.

IT (source); EN (target):



Conversational markers are natural in oral speech. Given that chat language is a mixture of written and spoken language, its transfer is expected. Conversational markers are linguistic structures that are very specific to their own language. Being so specific, it is preferable to tag them even when they are written correctly in the source text.

PT-BR (source):

(4a): [Ops]_{Conversational Marker}, falha ao tentar....tente mais tarde.

(4b): [Tchau]_{Conversational Marker} :)

(4c): [Maravilha]_{Conversational Marker}, obrigado pela ajuda.

In the example (4a), there is a conversational marker that was not translated and was kept the same in the target language, where it is written differently. While sometimes it is possible for two languages to share a conversational marker with graphical and purpose similarity, this example was not the case.

Source language	Source text	Linguistic structure	MT Target text	Post-edited Target text
PT_BR	Ops, falha ao tentar....tente mais tarde.	Ops	Ops, crash while trying.... try later.	Oops, it fails while trying...try later.

Table 21. Untranslated Portuguese Conversational Marker

The other two examples, (4b) and (4c), had conversational markers that were indeed translated, however they were badly translated, resulting in a different meaning in the target language.

Source language	Source text	Linguistic structure	MT Target text	Post-edited Target text
PT_BR	Tchau :)	Tchau	Wow:)	Bye :)
PT_BR	Maravilha, obrigado pela ajuda.	Maravilha	Wonder , thank you for the help.	That's wonderful , thank you for your help.

Table 22. Portuguese Conversational Marker translated incorrectly

Emoticons have become complementary to chat language. Users tend to show their emotions through emoticons, hence its name. This gives clues to what their moods currently are.

PT-BR (source):

(5a): É porque a aula está ocorrendo neste instante, não há essa opção de cancelamento [:/]Emoticon

(5b): Tchau [:)]Emoticon

(5c): Fale com uma pessoa [👤]Emoticon

The emoticons in the examples (5a) and (5b) had issues with whitespaces (with extra or missing whitespaces) in the target text. This resulted from the tokenization in “cancelation:” vs “cancelamento :”. When emoticons use the traditional punctuation markers, the tokenizer could cause its wrong translation.

Source language	Source text	Linguistic structure	MT Target text	Post-edited Target text
PT_BR	É porque a aula está ocorrendo neste instante, não há essa opção de cancelamento :/	:/	It is because the class is taking place right now, there is no such option of cancellation: /	It is because the class is taking place right now, there isn't that cancellation option :/
PT_BR	Tchau :)	:)	Wow:)	Bye :)

Table 23. Emoticons with whitespace issues

The example (5c) has an emoticon that changed its word order in the sentence. This could have been caused by the use of an “unicode” emoticon.





Source language	Source text	Linguistic structure	MT Target text	Post-edited Target text
PT_BR	Fale com uma pessoa 		Talk to a  person	Talk to a person 

Table 24. Emoticon with wrong word order

Customer Support deals with many emotions, especially negative ones. To express their frustration or anger, users resort to the use of profanities. Profanities are also linguistic structures that might have an impact on the target text. The only profanity found and annotated in the source was mostly translated in the target language, except for the word “luavas”. Although the piece of text that was translated with profanities in the target language, its syntax is very poor.

Source language	Source text	Linguistic structure	MT Target text	Post-edited Target text
RO	Raspundeti sclavilor luavas morți și copii in pula de jeguri ce sunteti	luavas morți și copii in pula de jeguri ce sunteti	Answer the dead luavas slaves and children in the cock of the shit you are	Answer, you slaves, you deserve I fuck your dead people and children, you human garbage.

Table 25. Profanity example

Neutral structures in *Agents'* data

In the *Agents'* data, the linguistic structures annotated were *Code Switching* (Table 26) and *Segmentation*. In *agents'* sources annotations, the source language was English and the target language German. Whenever the *Code Switching* issue type was used, the source text was already in German. For *agents* to use another language, besides the source language, it is usually a way to relate to the user by using their native language. The only example that was not correct in the target text had an abbreviation in English (“sry”) and its translation had a completely different meaning in the target language. And it also had an orthography error on the word “nivht” that should have been “nicht”.

Source language	Source text	Linguistic structure	MT Target text	Post-edited Target text
EN	sry es geht nivht	sry es geht nivht	Mach es kaputt nimm	Entschuldigung aber es geht nicht.

Table 26. *Agents'* data Code Switching example

There were no problems with emoticons, they were all correct in the target language. All the idiomatic expressions written by the *agents* were translated correctly in the target language. When it comes to *Segmentation*, *agents* have time restrictions where they are only given a couple of minutes or seconds to answer the user, which ultimately results in writing as

quickly as possible while troubleshooting all the user’s questions and thus segmenting the original message. There were a couple of segmented sentences in the agents’ data. In the first example there is a question that was divided into two chat messages.

EN (source); DE (target):

Source	MT Target	Post-edited Target text
Thank you, is this for the [PRODUCT]	Vielen Dank, ist das für die [PRODUCT]	Vielen Dank, ist das für die [PRODUCT] und die [PRODUCT]?
and the [PRODUCT]?	Und die [PRODUCT]?	

The other example is a sentence that begins with the conjunction “since” which introduces subordinate clauses and focuses mainly on the result of something. If the conjunction “since” implies a reason and a result, then the sentence should not be segmented into two different chat messages.

EN (source); DE (target):

Source	MT Target	Post-edited Target text
Since I don't want you to wait for too long.	Da ich nicht möchte, dass du zu lange wartest.	
I will process the refund now so you can receive it as soon as possible.	Ich werde die Rückerstattung jetzt bearbeiten, damit du sie so schnell wie möglich erhalten kannst.	Da ich nicht möchte, dass du zu lange wartest ich werde die Rückerstattung jetzt bearbeiten, damit du sie so schnell wie möglich erhalten kannst.

5.3.2 Critical errors analysis

Besides the impact of the *Neutral*, it was also important to check if there were any critical errors in the source text that had an impact on the target text. While there were not any critical errors in the agents' data, there were some critical errors annotated in the user's data. The examples below are all from the user's data.

RO (source):

(6a): [Brd]**Addition**

IT (source):

(6b): [Tutto il vuoi dei miei giorni]**Named Entity**

(6c): [Io sotto casa e non è nessuno]**Omission (X2)**

6d): [Tifare]**Lexical Selection** l'ordine

RO (source):

(6e): Banca trebuia sa [mii anuleze]**Word Order**

NL (source):

(6f): app is [bezit]**Lexical Selection**

As presented in examples from (6a) to (6f), the languages with critical errors on the source were Romanian, Italian and Dutch. These are the critical errors that had an impact on the target text. In this example, (6a), there was addition in the source. Being an addition meant that it should have been erased from the target text.

Source language	Source text	Issue type	Target text	Post-edited Target text
RO	Brd	Addition	br	[]

Table 27. Romanian critical Addition error

The example (6b) has a named entity, in this case a book title, that was written incorrectly. With this kind of error there would always be an error in the target text. However, since the book has not been translated into English, its title should be kept as its original one, in Italian, in the target text.

Source language	Source text	Issue type	Target text	Post-edited Target text
IT	Tutto il vuoi dei miei giorni	Named Entity	All you want of my days	Tutto il buio dei miei giorni

Table 28. Italian critical Named Entity error

Then, in the example (6c), there are two omissions in the source text that resulted in an incomprehensible target text.

Source language	Source text	Issue type	Target text	Post-edited Target text
IT	Io sotto casa e non è nessuno	Omission (X2)	I under the house and it is no one	I'm at your place and there's nobody home.

Table 29. Italian critical Omission error

The example below, (6d), was already explained concerning the *Segmentation* issue type, where the message had been fragmented. Instead of the verb “*tifare*”, which means “to cheer”, it should have been the verb “*rifare*” (“to redo” or “to replace”).

Source language	Source text	Issue type	Target text	Post-edited Target text
IT	Tifare l'ordine	Lexical Selection	Cheer the order	Replace the order.

Table 30. Italian critical Lexical Selection error

In the example (6e), there was an issue of *Word Order* in the source that turned into an *Omission* of “a few” in the target text.

Source language	Source text	Issue type	Target text	Post-edited Target text
RO	Banca trebuia sa mii anuleze	Word Order	The bank had to cancel thousands	The bank had to cancel a few thousands.

Table 31. Romanian critical Word Order error

Finally, the example (6f) had a *Lexical Selection* issue where instead of “*bezit*” (“property”), it should have been “*bezet*” (“busy”/“occupied”). This word changes the entire meaning of the sentence, resulting in a strange expression in the target text.

Source language	Source text	Issue type	Target text	Post-edited Target text
NL	app is bezit	Lexical Selection	App is owned	The app is running.

Table 32. Dutch critical Lexical Selection error

5.3.3 Typology Misusage

While analyzing the data, it was possible to see that some of the issue types that fall on the *Neutral* severity were not annotated. This is something that could be expected due to being the first pilot with the Source Typology and it takes practice to annotate according to its guidelines. The issue type that stood out the most in not being annotated was *Emoticon*. The reason for this is because emoticons have become so common and natural in chat language that we no longer look at them as an addition to the text, but rather as a part of it. In a total of 44 emoticons that were not annotated, only 10 had an impact on the target text. In the

following tables, we present the examples in both user and agent data, respectively, that had an impact on the target text.

Emoticons not annotated in *User data*

- (7a): Obrigado por escrever 😊
- (7b): Fico feliz em ajudá-lo hoje ☀️ &@@@
- (7c): Hi, we can write in English :)
- (7d): Done :)
- (7e): Obrigado por escrever 😊
- (7f): desculpe so mais uma pergunta :)
- (7g): danke :)
- (7h): Necesito una solución urgente:(
- (7i): Helaas krijg weer de foutmelding 😞😞

In the user data, the emoticons had different outcomes in the target text. In the examples (7a) and (7e), the emoticons interfered with the translation of the target text by considering the emoticon a word.

Source language	Source text	Emoticon	MT Target	Post-edited Target text
PT_BR	Obrigado por escrever 😊	😊	Thanks for writing in 😊	Thanks for writing 😊
PT_BR	Obrigado por escrever 😊	😊	Thanks for writing in 😊	Thanks for writing 😊

Table 33. Non-annotated emoticons that had impact on the translation

In the examples (7c), (7d), (7f) and (7h) had problems with whitespaces due to the use of traditional punctuation marks.

Source language	Source text	Emoticon	MT Target	Post-edited Target text
DE	Hi, we can write in English :)	:)	Hi, we can write in English:)	Hi, we can write in English :)
DE	Done :)	:)	Done:)	Done :)
PT_BR	desculpe so mais uma pergunta :)	:)	Sorry only one more question:)	Sorry only one more question :)
ES	Necesito una solución urgente:(:(I need an urgent solution: (I need an urgent solution :(

Table 34. Non-annotated emoticons with whitespace issues

In the example (7i) there was a minor *Word Order* error in the target text caused by the “unicode” emoticon.

Source language	Source text	Emoticon	MT Target	Post-edited Target text
NL	Helaas krijg weer de foutmelding 😞😞	😞😞	Unfortunately, get the error message 😞😞 again	Unfortunately, I get the error message again 😞😞

Table 35. Non-annotated emoticon with Word Order error

The machine changed the form of the emoticon in (7b), however the meaning was kept the same.

Source language	Source text	Emoticon	MT Target	Post-edited Target text
PT_BR	Fico feliz em ajudá-lo hoje ☀️ &@@@	☀️	Happy to help you out today 🌞	Happy to help you out today 🌱

Table 36. Non-annotated emoticon with different form

The emoticon in (7g) was omitted by the machine in the target text.

Source language	Source text	Emoticon	MT Target	Post-edited Target text
DE	danke :)	:)	Thank you []	Thank you :)

Table 37. Omitted non-annotated emoticon

Emoticons not annotated in *Agent* data

Source language	Source text	Emoticon	MT Target	Post-edited Target text
EN	I also like it, thank you:)	:)	Mir geht es auch gut, danke :)	Das mag ich auch, danke :)

Table 38. Internal pilot non-annotated emoticon in *Agent* data

In agents data, there was only an emoticon that was not annotated. By not having the correct whitespace in the source text, it interfered with its translation of the target text.

Conversational Markers not annotated

Source language	Source text	Conversational Marker	Target text
PT_BR	hahahhaaha	hahahhaaha	hahahhaaha
PT_BR	Bom, agora sim ficou bem claro	Bom	Well, now yes it was clear
DE	ugh	ugh	ugh
DE	Oh!	Oh	Oh.
NL	De foutmelding is: Oeps!	Oeps	Oops

Table 39. Internal pilot non-annotated Conversational Markers

While emoticons were fairly easy to check if they were annotated or not, conversational markers were more difficult. Although some languages have conversational markers in common, most of them are very unique to their language making it only possible for native speakers or long-term learners of a language to know them. The ones that were captured despite this were all well translated in the target text.

5.3.4 Inter Annotator Agreement

Since this case study was annotated in multiple languages by multiple annotators, we decided to see the results of Inter Annotator Agreement (IAA) and its impact. The IAA is a common practice performed after the annotation process and it is used for multiple purposes, such as “validating and improving annotation schemes and guidelines, identifying ambiguities or difficulties in the source, or assessing the range of valid interpretations” (Artstein, 2017:298). Most importantly, the IAA allows us to see and evaluate the reliability of the annotation process. That reliability will then allow us to see what needs to be improved

or what is working in our annotation guidelines. Artstein (2017) sums up this process perfectly with a simple figure, as shown in *Figure 23*.

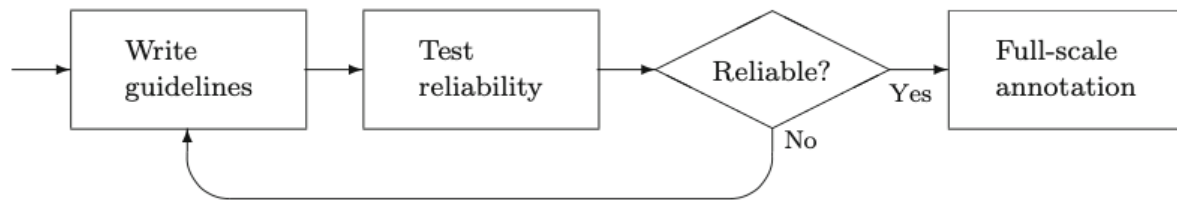


Figure 23. Iterative reliability testing (Artstein, 2017:299)

Through the IAA, we can see the effectiveness of the Source Typology and its guidelines. In the making of the typology, several issue types were joined together in order to ease the annotation process, while in the annotation guidelines we presented a section with tricky cases that might come up during the annotation process and also provided two decision trees. However, there are still some issue types that might have brought confusion to the annotators. Given that the annotation process is not possible to be performed mechanically but rather through human judgments, it is expected that there is going to be a variation in the agreement, not only with other annotators but also with themselves. (Artstein, 2017). As Lommel *et al.* (2014) points out, there is another aspect that needs to be taken into account, which is annotators' personal opinion. Despite providing guidelines and thorough explanations of issue types, annotators might disagree with them and have different definitions of what an error is. Although a minor aspect, this will highly impact the IAA. It is possible to measure the IAA with several coefficients. In this case study, we will only focus on two of them — the Cohen's Kappa coefficient and the Pearson product-moment correlation coefficient. Coefficients like Kappa "are intended to calculate the amount of agreement that was attained above the level expected by chance or arbitrary coding." (Artstein, 2017:300). More specifically, Cohen's Kappa is a quantitative coefficient that measures the agreement between two raters, in this case annotators, that are rating the same content. Kappa's values can be a negative number (less than 0), 0 and 1. A negative value means that there is almost no agreement, 0 means there is a random agreement between the annotators and 1 means that there is a complete agreement between both sides.

The Pearson correlation coefficient, represented by r , is “a measure of the strength of a linear association between two variables” (Laerd statistics, 2018). Its values range from -1 to 1. If the value is 0, there is no association between the two variables. So, if the value is higher than 0, there is a positive association between both variables, meaning that as the value of one variable increases so does the other. If the value is lower than 0, there is a negative association between the variables, where the value of one variable increases while the other decreases (Laerd statistics, 2018). While agreement coefficients are used for the improvement of annotation guidelines and “for data analysis to give a picture of how distant the annotators’ interpretation of the phenomena is” (Amidei *et al.*, 2019:352), correlation coefficients indicate “to what extent annotators are consistent with each other” (Amidei *et al.*, 2019:352).

5.3.4.1 Study of IAA of Internal pilot

Having this in mind, we will present the results of the IAA in the internal pilot. Due to time constraints, it was not possible to perform IAA on the outbound data. For that reason, the IAA was only performed with the User’s data. In order to identify the annotator and the language, we will present the annotator with the corresponding language code. Firstly, we will analyze the intra-annotator agreement results and then the results of the inter-annotator agreement. Analyzing the intra-annotator agreement is important because it will allow us to see if the annotators had any difficulties with the learning of the typology and, as Artstein states, making more detailed analysis “can uncover unreliable facets of an otherwise reliable annotation process” (Artstein, 2017:310). Since we are dealing with user’s data, it is common to have repeated segments. Having consistency with themselves or not in the annotation process is an indicator of the typology, whether it is easy to understand or not. We will start with the Pearson values. In *Table 40*, we will present the agreement on a segmental level.

Annotator	Segment-level intra-agreement value
PL	0.9
ES-1	1.0
ES-2	0.7
RO	0.9
NL	1.0
PT-BR	0.9
IT-1	1.0
IT-2	0.9
DE-1	1.0
DE-2	1.0

Table 40. Intra-annotator agreement Pearson values on a segmental level

As can be seen in *Table 40*, all the annotators were consistent with themselves with values of 1 and 0.9, except for ES-2 with a value of 0.7. Artstein explains that the computational linguistic community has accepted Carletta’s recommendation with “accepting coefficient values above 0.8 as reliable, with somewhat lower values also considered acceptable in certain circumstances” (Artstein, 2017:302). In Amidei *et al.* (2019), by gathering several research papers on IAA, it was possible to present the average, minimal and maximum IAA values of different coefficients. The minimal value for Pearson’s r is 0.20. The results presented in *Table 40* were farther from that value, which shows that, overall, the Source Typology is being efficient and its annotation guidelines were simple and explicit for the annotators to understand how this typology works. The lower value is from the second annotator for Spanish. Although this value is not necessarily low in terms of having no inner consistency, it is still low compared to the values from the other annotators.

Despite being consistent with themselves, that does not necessarily mean that the annotators with the same language pair were consistent with each other. Now, we will present the Pearson values of the inter-annotator agreement on a segmental and document level. For that reason, the only languages shown will be Spanish, Italian and German.

Language	Segment-level IAA value
ES	0.5
IT	0.5
DE	0.1

Table 41. Inter-annotator agreement Pearson values on a segmental level

Language	Document-level IAA value
ES	0.1
IT	0.5
DE	-0.1

Table 42. Inter-annotator agreement Pearson values on a document-level

From both tables 41 and 42, we can see that the IAA was positive on a segmental level, despite not being very high, especially in the case of the agreement of the annotators for German with the Pearson value of 0.1. On a document-level, only the annotators for Italian kept their consistency with a 0.5 value, while the agreement values of the other annotators decreased. Although the value is not very high, the agreement between the annotators for Spanish was still positive. However, the agreement between the annotators for German was negative (-0.1). These results were very general, so we used the Kappa coefficient to show the agreement in categories, issue types and severities.

Language	Categories IAA value
ES	0.3
IT	0.2
DE	0.2

Table 43. Categories Inter-annotator agreement Kappa values

Language	Issue types IAA value
ES	0.3
IT	0.3
DE	0.2

Table 44. Issue types Inter-annotator agreement Kappa values

Language	Severities IAA value
ES	0.3
IT	0.2
DE	0.2

Table 45. Severities Inter-annotator agreement values

As mentioned before Amidei *et al.* (2019), several research papers on IAA were gathered, making it possible to present the average, minimal and maximum IAA values of different coefficients. The minimal value for Cohen's Kappa is 0.10. Although the values are not lower than 0, which would mean that there was not any agreement between the annotators, the values are still higher than the minimal value. Values around 0.2 and 0.3 means that the agreement is somewhat random between the annotators. In order to illustrate these results, some examples of each language will be displayed.

È già vicino al router, comunque il problema lo ho riscontrato dopo più di un giorno di utilizzo normale quindi se il problema fosse stato quello si sarebbe dovuto presentare fin da subito

È già vicino al router, comunque il problema lo ho riscontrato dopo più di un giorno di utilizzo normale quindi se il problema fosse stato quello si sarebbe dovuto presentare fin da subito

Figure 24. Italian IAA

also , meinem lapotop NAMED ENTITY
 funktioniert ORTHOGRAPHY nichts ORTHOGRAPHY mehr richtig WHITESPACE , PUNCTUATION
 wenn OMISSION den anmache LEXICAL SELECTION PUNCTUATION , OMISSION
 kommt LEXICAL SELECTION OMISSION normalerweise Abgabe von Passwort oder
 kennwort LEXICAL SELECTION PUNCTUATION , aber jetzt das bilschirm WORD ORDER
 ist WORD ORDER lehr ORTHOGRAPHY und man kann ncihts ORTHOGRAPHY dazu machen

also CAPITALIZATION WHITESPACE , meinem lapotop [] fonktiomert nichts mehr richtig
 WHITESPACE , wenn den anmache WHITESPACE , kommt normalerweise Abgabe von
 Passwort oder kennwort WHITESPACE , aber jetzt das bilschirm ist lehr und man kann
 ncihts dazu machen

Figure 25. German IAA

Pude solucionar lo del microfono pero lo de el 7.1 no, escucho todo muy bajo y yo siempre que ocupaba el 7.1 se escuchaba bien todo y fuerte.

Pude solucionar lo del microfono ORTHOGRAPHY pero lo de el WRONG FUNCTION WORD 7.1 no,
 escucho todo muy bajo y yo ADDITION siempre que ocupaba el LEXICAL SELECTION 7.1 se
 escuchaba bien todo PUNCTUATION y fuerte.

Figure 26. Spanish IAA

With these examples, we can understand the Kappa values better. In *Figure 24*, while one annotator decided to annotate three errors in a sentence, the other annotator only detected one error and it was not even in agreement with the other annotator. In the following example, the same happens where one annotator detects more than the other without having any agreement with each other, despite one whitespace error. In the final example, we can see that there was a complete disagreement between the annotators for Spanish. While one felt that there were several errors to be annotated, the other annotator did not detect any error in the same sentence.

In conclusion, the IAA results for inter and intra annotator agreement had very satisfactory results. The agreement was tested at different levels, segment and document, and with different factors, categories, issue types and severities and most of its results had a positive value. This proves that our typology and guidelines were effective and reliable to our annotators.

6. Language eGuides

While working on the Source Typology, another project was proposed by the Product Marketing team at Unbabel. This project involved producing Language eGuides. The purpose of the Language eGuides was to provide key points when writing in English for agents that were not English native speakers and guide them in the cultural context of their clients. With this supporting material, agents would be able to provide a good English input so that the message was appropriate for the target audience. Given that every culture has different ways of communicating that go beyond writing, it was important to provide a cultural context of the target language of their output. For instance, in some languages customer support is carried out in a very formal register, while in other languages it is more common to communicate in an informal tone. Besides this, it was also important to have machine-friendly English content. In total, 17 Language eGuides were written. The eGuides were provided in German, French, Dutch, Italian, European Portuguese, Brazilian Portuguese, European Spanish, Latin American Spanish, Danish, Norwegian, Swedish, Finnish, Japanese, Indonesian, Korean, Vietnamese, Simplified Chinese and Traditional Chinese (joined). Although the focus was mostly on European languages, there was an effort to include Asian languages. These eGuides were then shared with Unbabel clients.

In order to write these Language eGuides, it was important to analyze the source data from the agents. In this case, both chat and tickets data was analyzed. This made it possible to see which errors were most common from the agents and what needed to be improved on their side. The majority of errors were typography errors, such as *Punctuation*, *Capitalization* and *Whitespace*. As the agent annotation performed in the beginning of creating the Source Typology, there were many errors that only happened due to English not being their first or second language, such as *Omission* and *Word Order*. These errors not only affected the input, but also the output. For that reason, it was necessary to also check the output and that way understand the impact of the errors on MT. Through this analysis, it was possible to provide tips on how to write in English so that the MT output was optimized.

Firstly, these Language eGuides provided general English writing tips and ended on a more specific note according to the target language. To gather some information on English writing tips, a guide provided by the European Commission was consulted, “How to write

clearly”²⁷. This guide is available in all EU official languages. From that, it was possible to summarize on how to write in a clear and simple way in the English language, such as avoiding writing long sentences or using the passive voice. Analyzing the source from the agents also allowed us to write some tips concerned more on being machine-friendly, where typography errors could become problematic, and abbreviations and idiomatic expressions might not be captured by the machine and be translated incorrectly. For example, there are some English expressions (‘Hello, John here.’; ‘I hope this email finds you well’) that are usually used in customer support, but they do not translate well in the other languages. In all writing tips, it was also provided an alternative or solution in order to guide agents, in this case the alternatives were ‘Hello, I’m John.’ and ‘I hope you’re well.’.

The final section of the Language eGuides was more specific in terms of the cultural context of the target language. To start this section, we provided the rules of greetings and closings that are suitable for that particular language. This information was available in some Unbabel Language Guidelines²⁸. These Guidelines are very complete and helped to see the differences in the greetings and closings of each language, for example which register was required and how the punctuation was used in them. Besides this research, whenever it was possible, we tried to talk to someone native of that language to provide more information about more specific cultural aspects. One of the perks of having a multicultural company is being able to ask for help from an Unbabel employee. If that was not possible, then we also asked annotators, who were natives or proficient of those languages, to provide a cultural insight. Each Language eGuide is unique according to specificities of the language required.

Writing these eGuides was essential for the work on the Source Typology. By analyzing the source from the agents, it was possible to check which errors were most common and what issue types were used to describe them. This analysis also allowed us to see if there were any errors that were not accounted for but still had an impact on the MT, such as the use of abbreviations and emojis. The information gathered from this analysis was also ultimately used as base for the new additions on the Source Typology.

²⁷ <https://op.europa.eu/en/publication-detail/-/publication/c2dab20c-0414-408d-87b5-dd3c6e5dd9a5>

²⁸ The Unbabel Language Guidelines are guides for the post-editors and they provide the grammar rules and localization challenges that might come up during the process of post-edition.

7. Conclusions and Future work

This dissertation aimed to prove the propagation of source errors in the TT and further investigate the reasons behind those errors.

Firstly, it was important to understand and learn all the previous work done on source errors and the typologies available. Secondly, it was crucial to be fully aware of the customer service environment and understand the linguistic challenges it may bring. As both agents and users have to deal with stressful situations, it is expected to not have a perfect source text. Even the content type of chat brings out many specific challenges, such as immediacy and the mixture between spoken and written language. Thirdly, one great motivation for this project was to analyze real data and have ecological examples of source errors. So, building a data driven typology allowed us to verify what kind of errors were being made and the reasons behind them.

In our first annotation effort, 44,302 words from the agent and user were annotated and this made it possible for us to then verify which errors occurred at a monolingual level and which ones were missing from the typologies previously examined. With these results, we could then build the Source Typology and begin testing it.

We conducted an internal pilot with three case studies: *PT-BR inbounds*, *Agent Annotation* and *Multilingual internal pilot*. In total, in this internal pilot 26,855 words, 2802 source errors and 239 neutral severity linguistic structures (e.g., discourse markers, emojis) were annotated.

In the first study case, *PT-BR_EN inbounds*, the source language was Brazilian Portuguese and the target language was English. The source MQM was 72.26 and 2909 words were annotated. A comparison was made with the ST and TT, so it was decided to test different MT systems and annotate them with the Unbabel Error Typology. In total, 8876 words of the TT were annotated. The MQM of the TT was much lower than the source one, where the lowest result was 29.05 and the highest was 51.37. The reason behind these results was that while the ST mainly had *Minor* errors, such as *Orthography*, *Diacritics* and *Punctuation*, the TT had a considerable amount of *Major* and *Critical* errors. With this information, we aligned source and target, checking the same errors found in both, source errors that originated different errors in the TT and the neutral structures that had an impact on the MT output. 34 source errors were transferred to the TT, 29 source errors originated

different errors in the target, where *Typography* errors created critical *Lexical Selection* and *Unintelligible* errors, and 9 neutral structures had an impact on the TT, mostly abbreviations used in the source that created critical *Unintelligible* errors.

In the second study case, *Agent Annotation*, the source language was English and the target language was French. In this study, the opposite occurred. Although the source MQM was already high (78.53), the target MQM was even higher (87.41). In total, 20,555 words were annotated in this study case. These results occurred due to the TT having less *Minor* and *Major* errors. The alignment between both texts was also performed, where 59 source errors were transmitted to the TT, 40 errors in the ST originated different errors in the TT, whose more serious example was when *Orthography* errors created critical *Lexical Selection* and *Untranslated* errors, and no neutral structures had an impact on the TT.

In the third study case, *Multilingual internal pilot*, multiple languages were annotated in both directions. The users' data annotated was in Dutch, Polish, Romanian, Brazilian Portuguese, Italian, Spanish and German and the agents' data annotated was in English. In the users' data, the lowest MQM was the German source (53.09) and the highest MQM was the Dutch source (90.88), and in the agents' data the MQM was 85.46. In this study, in total, 14,098 words were annotated. Given the multiplicity of languages and data, we have looked into the neutral structures that had an impact on the MT output, the *Critical* errors found in the ST and the typology misuse. In the users' data, a lot of different neutral structures had impact in the MT, namely *Code Switching*, *Idiomatic*, *Abbreviation*, *Segmentation*, *Conversational Marker*, *Emoticon* and *Profanity*, and there were several *Critical* errors, annotated with the *Addition*, *Omission*, *Named Entity*, *Lexical Selection* and *Word Order* issue types. In the agents' data, only two neutral structures had an impact on the TT, *Code Switching* and *Segmentation*, and there were no *Critical* errors identified. During the analysis of the neutral structures, we could observe that some of those structures were not annotated by our participants in the pilot. The most evident were emoticons and conversational markers. The reason pointed out by our participants for this was because these specific structures are so common in online conversations that we tend to overlook them.

After performing the internal pilot, it was important to have an unbiased evaluation of the effectiveness and reliability of the Source Typology. For that purpose, we performed IAA scores. Through agreement coefficients it was possible to see the consistency between the annotators. Besides that, it was also performed intra annotator agreement, which is the

consistency that the annotators have with themselves. This allowed us to understand the difficulty presented in the typology. The IAA was performed on two levels, segment and document level, and took into account different agreement factors (categories, issue types and severities). Given that the results were mainly positive, we can prove that our typology is reliable and effective for our annotators. Although further analysis of the IAA, tackling mostly potential ambiguous categories, may provide better insights in this topic.

In conclusion, it was proven that source errors have an actual effect on the TT. Although *Critical* and *Major* errors can have an impact on MT output, we could verify that even *Minor* errors or *Neutral* structures in the ST can create *Critical* errors in the TT. The version of the Source Typology presented here is just a prototype.

Currently, the Source Typology is being tested even further in order to be aligned with the new version of the Unbabel Error Typology and with the end goal of having an harmonized Unbabel Error Typology that takes into account source and target errors. The main goal is to have this typology ready for production and then provide it as an Unbabel service. Our work is already contributing to a research project called MAIA²⁹ and to other Master theses on automatic source errors identification. In addition, the analytics conducted in this dissertation is already being used as a source of information also within some teams at Unbabel.

²⁹ <https://aclanthology.org/2020.eamt-1.68.pdf>

Bibliography

- Alcina, A. (2017, December). Entrevista amb Mikel Forcada. *Revista Tradumàtica*.
https://revistes.uab.cat/tradumatica/article/view/n15-alcina-forcada/pdf_45
- 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://www.aclweb.org/anthology/W19-8642.pdf>
- Artstein, R. (2017). *Handbook of Linguistic Annotation* (N. Ide & J. Pustejovsky, Eds.). Springer. 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.
<https://www.cs.cmu.edu/~alavie/METEOR/pdf/Banerjee-Lavie-2005-METEOR.pdf>
- Bar-Hillel, Y. (1953). Some Linguistic Problems Connected With Machine Translation. *20(Philosophy of Science)*, 217-225.
- Bar-Hillel, Y. (1960). The Present Status of Automatic Translation of Languages. *1(Advances in Computers)*, 91-163.
https://cesaraguilar.weebly.com/uploads/2/7/7/5/2775690/bar_hillel_01.pdf
- Bentivogli, L., Forner, P., & Pianta, E. (2004, August 23). Evaluating Cross-Language Annotation Transfer in the MultiSemCor Corpus.
<https://www.aclweb.org/anthology/C04-1053.pdf>
- Bishop, J. (2021, January 13). *Idiom Examples That Break the Grammar Mold - BKA Content*. BKA Content - Buy Articles, Blog Posts, Website Content and More!

- Retrieved February, 2021, from
<https://www.bkacontent.com/idiom-examples-that-break-the-grammar-mold/>
- Brussel, L. V., Tezcan, A., & Macken, L. (2018). A Fine-grained Error Analysis of NMT, PBMT and RBMT Output for English-to-Dutch. 3799-3804.
<https://core.ac.uk/download/pdf/158345516.pdf>
- Buchicchio, M. (2017). Português controlado para a tradução automática: Português-Italiano.
<https://repositorio.ul.pt/handle/10451/28715>. Retrieved September, 2020, from
https://repositorio.ul.pt/bitstream/10451/28715/1/ulfl233817_tm.pdf
- Burchardt, A., & Lommel, A. (2014, November 19). Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality. 3-11.
<http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf>
- Burchardt, A., Lommel, A., & Macketanz, V. (2020). A new deal for translation quality.
<https://link.springer.com/article/10.1007%2Fs10209-020-00736-5>
- Cabarrão, V., Moniz, H., Batista, F., Ferreira, J., Trancoso, I., & Mata, A. I. (2018, June). Cross-domain analysis of discourse markers in European Portuguese (A. Stent, Ed.). 79-106. 10.5087/dad.2018.103
- Chang, C. B., & Mishler, A. (2012). Evidence for language transfer leading to a perceptual advantage for non-native listeners. 2700-2710. <http://dx.doi.org/10.1121/1.4747615>
- Comparin, L., & Mendes, S. (2017). Using error annotation to evaluate machine translation and human post-editing in a business environment.
<https://www.semanticscholar.org/paper/Using-error-annotation-to-evaluate-machine-and-in-a-Comparin/c9d28db57b3cedfd75a2fe694dcc59ba8caf7029?p2df>
- Directorate-General for Translation (European Commission). (2011, March 16). *How to write clearly*. Publications Office of the EU. Retrieved Setembro, 2020, from

<https://op.europa.eu/en/publication-detail/-/publication/c2dab20c-0414-408d-87b5-dd3c6e5dd9a5>

EUR-Lex. (2016, April 27). *Official Journal of the European Union*. EU law - EUR-Lex.

Retrieved September, 2020, from

<https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN>

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021, April 29).

Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. 1-22. <https://arxiv.org/pdf/2104.14478.pdf>

Galvao, G. C.T. (2009, May 27). Linguistic interference in translated academic texts: A case study of Portuguese interference in abstracts translated into English. 1-25.

<https://www.diva-portal.org/smash/get/diva2:221865/FULLTEXT01.pdf>

Gamon, M. (2010, January). Using Mostly Native Data to Correct Errors in Learners'

Writing: A Meta-Classifer Approach. *ResearchGate*. Retrieved August, 2021, from

https://www.researchgate.net/publication/228937862_Using_mostly_native_data_to_correct_errors_in_learners'_writing_A_meta-classifier_approach

German Research Center for Artificial Intelligence (DFKI) & QTLaunchPad. (2015,

December 30). *Multidimensional Quality Metrics (MQM) Definition*. Retrieved

November, 2020, from

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

Gilmartin, E., Vogel, C., & Campbell, N. (2017, August 18-19). Disfluency in chat and chunk phases of multiparty casual talk. *Proceedings of DiSS 2017*, 25-28.

https://www.ida.liu.se/~robek28/conferences/diss2017/DiSS2017_Gilmartin_Vogel_Campbell.pdf

- Golonka, E. M., Tare, M., & Bonilla, C. (2017, June). Peer interaction in text chat: Qualitative analysis of chat transcripts. *21*, 157-178.
https://scholarspace.manoa.hawaii.edu/bitstream/10125/44616/1/21_02_golonkatarebonilla.pdf
- Guy, G. R., & Zilles, A. M.S. (2006). Endangered Language Varieties: Vernacular Speech and Linguistic Standardization in Brazilian Portuguese. 1-21.
<http://gregoryguy.com/wp-content/uploads/Guy-Zilles-2006-Endangered-Language-Varieties-Vernacular-Speech-and-Linguistic-Standardization-in-Brazilian-Portuguese-GURT.pdf>
- Hammarberg, B., & Grigonytė, G. (2014). Non-Native Writers' Errors – a Challenge to a Spell-Checker.
<https://www.diva-portal.org/smash/get/diva2:764515/FULLTEXT01.pdf>
- Herzig, J., Feigenblat, G., Shmueli-Scheuer, M., Konopnicki, D., Rafaeli, A., Altman, D., & Spivak, D. (2016, September). Classifying Emotions in Customer Support Dialogues in Social Media. *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 64–73.
<https://www.aclweb.org/anthology/W16-3609.pdf>
- Hohn, S., Pfeiffer, A., & Ras, E. (2016). Challenges of error annotation in native/non-native speaker chat. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 114-124.
https://linguistics.rub.de/konvens16/pub/15_konvensproc.pdf
- Hutchins, J. (1996, June). ALPAC: the (in)famous report. (MT News International), 9-12.

- Hutchins, W. J. (1995). Machine Translation: a Brief History. (Concise history of the language sciences: from the Sumerians to the cognitivists), 431-445.
https://www.infoamerica.org/documentos_pdf/bar05.pdf
- Hutchins, W. J. (2001). Machine translation over fifty years. (Histoire, Epistemologie, Langage), 7-31.
<http://paginaspersonales.deusto.es/abaitua/konzeptu/ta/hutchins01.htm>
- Hutchins, W. J. (2003). Machine translation: a concise history. 1-21.
<https://opencourses.ionio.gr/modules/document/file.php/DFLTI199/%CE%95%CE%B2%CE%B4%CE%BF%CE%BC%CE%AC%CE%B4%CE%B1%204/Machine%20Translation%20-a%20concise%20history.pdf>
- Kato, M. A., & Martins, A. M. (2016). The Main Varieties of Portuguese: an overview on word order. 1-29. <https://repositorio.ul.pt/bitstream/10451/31197/1/Martins2016a.pdf>
- Kenny, D. (2018). *The Routledge Handbook of Translation and Philosophy*. Routledge.
- Kepler, F., Trénous, J., Treviso, M., Vera, M., & Martins, A. F.T. (2019). OpenKiwi: An Open Source Framework for Quality Estimation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 117-122.
 10.18653/v1/P19-3020
- Koehn, P. (2020). *Neural Machine Translation* (The Johns Hopkins University ed.). Cambridge University Press. 10.1017/9781108608480
- Kraichoke, C. (2017, May). Error Analysis: A Case Study on Non-Native English Speaking College Applicants' Electronic Mail Communications. *ShocalWorks@UARK*. Retrieved August, 2021, from <https://scholarworks.uark.edu/etd/1910/>
- Laerd statistics. (2018). *Pearson Product-Moment Correlation*. SPSS Statistics Tutorials and Statistical Guides. Retrieved May 24, 2021, from

- <https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php>
- Lee, J., & Seneff, S. (2009, January). An analysis of grammatical errors in non-native speech in English. *ResearchGate*. 10.1109/SLT.2008.4777847
- Lind, A. (2012, June 12). Chat Language - In the continuum of speech and writing. 1-25.
<https://www.diva-portal.org/smash/get/diva2:548889/FULLTEXT01.pdf>
- Lommel, A. (2019, April 11). *MQM Top Level (2019-04-11)*. W3C Community and Business Groups. Retrieved November, 2020, from
<https://www.w3.org/community/mqmcg/mqm-top-level-2019-04-11/>
- Lommel, A., Popović, M., & Burchardt, A. (2014, May). Assessing Inter-Annotator Agreement for Translation Error Annotation. *Conference: LREC Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*.
- Lommel, A., Uszkoreit, H., & Burchardt, A. (2014, December). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista tradumàtica*.
https://www.researchgate.net/publication/307174965_Multidimensional_Quality_Metrics_MQM_A_Framework_for_Declaring_and_Describing_Translation_Quality_Metrics
- Lopes, R. T. d. S. (2019). Qualidade na tradução automática e na pós-edição: anotação de erros de concordância e ordem de palavras.
<https://repositorio.ul.pt/handle/10451/41784>. Retrieved April, 2021, from
https://repositorio.ul.pt/bitstream/10451/41784/1/ulfl274984_tm.pdf
- Lüdeling, A., & Hirschmann, H. (2015, November 6). Error annotation systems. 136-156.
https://www.researchgate.net/profile/Hagen-Hirschmann/publication/291835319_Error

r_annotation_systems/links/571dfbf408aead26e71a7128/Error-annotation-systems.pdf?origin=publication_detail

Mattiello, E. (2013). *Extra Grammatical Morphology in English - Abbreviations, Blends, Reduplicatives and Related Phenomena* (E. C. Traugott & B. Kortmann, Eds.; Topics in English Linguistics 82 ed.). De Gruyter Mouton.

<https://www.degruyter.com/document/doi/10.1515/9783110295399/html>

Mercader-Alarcón, J., & Sánchez-Martínez, F. (2016, December). Analysis of translation errors and evaluation of preediting rules for the translation of English news texts into Spanish with Lucy LT. *Revista Tradumàtica: tecnologies de la traducció*.

<https://www.dlsi.ua.es/~fsanchez/pub/pdf/mercader-alarcon16.pdf>

Miyata, R., & Fujita, A. (2021, February 5). Understanding Pre-Editing for Black-Box Neural Machine Translation. *arXiv:2102.02955*. Retrieved February, 2021, from

<https://arxiv.org/pdf/2102.02955.pdf>

Moorkens, J., Castilho, S., Gaspari, F., & Doherty, S. (Eds.). (2018). *Translation Quality Assessment: From Principles to Practice* (Machine Translation: Technologies and Applications ed., Vol. 1). Springer. <https://doi.org/10.1007/978-3-319-91241-7>

Nasr, A., Damnati, G., Guerraz, A., & Bechet, F. (2016, September 13-15). Syntactic parsing of chat language in contact center conversation corpus. (Proceedings of the SIGDIAL 2016 Conference), 175-184. <https://www.aclweb.org/anthology/W16-3621.pdf>

Oliveira, L. d. S. d. N. (2017). Expressões Fixas do Português Formadas a partir de Nomes Gerais: aspectos lexicais e variacionistas. 9-80.

<http://www.poslin.letas.ufmg.br/defesas/1804M.pdf>

- Otemuyiwa, A. A. (2017, June 8). A Linguistic Analysis of WhatsApp Conversations among Undergraduate Students of Joseph Ayo Babalola University. 5, 393-405.
<https://doi.org/10.22158/selt.v5n3p393>
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002, July). BLEU: a Method for Automatic Evaluation of Machine Translation. (Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)), 311-318.
<https://www.aclweb.org/anthology/P02-1040.pdf>
- Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020, October 19). COMET: A Neural Framework for MT Evaluation. <https://arxiv.org/pdf/2009.09025.pdf>
- Roturier, J., & Bensadoun, A. (2011). Evaluation of MT Systems to Translate User Generated Content. 244-251.
<https://www.semanticscholar.org/paper/Evaluation-of-MT-Systems-to-Translate-User-Content-Roturier-Bensadoun/8763d41f730dcaf11fec1189390a66ac32e66964>
- Rozovskaya, A., & Roth, D. (2010, September). Annotating ESL Errors: Challenges and Rewards. *ResearchGate*. Retrieved August, 2021, from
https://www.researchgate.net/publication/228817793_Annotating_ESL_Errors_Challenges_and_Rewards
- Rubino, R. B., & Pine, J. M. (2018, February). Subject–verb agreement in Brazilian Portuguese: what low error rates hide. *Journal of Child Language*.
https://www.researchgate.net/publication/13678788_Subject-verb_agreement_in_Brazilian_Portuguese
- Ruzaitė, J., Dereškevičiūtė, S., Kavaliauskaitė-Vilkinienė, V., & Krivickaitė, E. (2020, September). Error Tagging in the Lithuanian Learner Corpus. *ResearchGate*. 10.3233/FAIA200631

Salesforce. *What is CRM? - Salesforce EMEA*. CRM Software & Cloud Computing Solutions - Salesforce EMEA. Retrieved September, 2020, from

<https://www.salesforce.com/eu/learning-centre/crm/what-is-crm/>

Sanchez-Torron, M., & Koehn, P. (2016). Machine Translation Quality and Post-Editor Productivity (S. Green & L. Schwartz, Eds.). *Volume 1: MT Researchers' Track* (The Twelfth Conference of The Association for Machine Translation in the Americas), 16-26.

https://amtaweb.org/wp-content/uploads/2016/10/AMTA2016_Research_Proceedings_v7.pdf

Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning Robust Metrics for Text Generation. <https://www.aclweb.org/anthology/2020.acl-main.704.pdf>

Shei, C.-C. (2002, January). Teaching MT Through Pre-editing: Three Case Studies.

<https://aclanthology.org/2002.eamt-1.10.pdf>

Sinha, A., Banerjee, N., Sinha, A., & Shastri, R. K. (2009, September). Interference of first language in the acquisition of second language. *1(7)*(Journal of Psychology and Counseling), 117-122.

https://academicjournals.org/article/article1379761693_Sinha%20et%20al.pdf

Sirbu, A. (2015, May). Language Interference Triggered by Bilingualism. *"Mircea cel Batran" Naval Academy Press*.

https://www.researchgate.net/publication/337472381_LANGUAGE_INTERFERENC_E_TRIGGERED_BY_BILINGUALISM

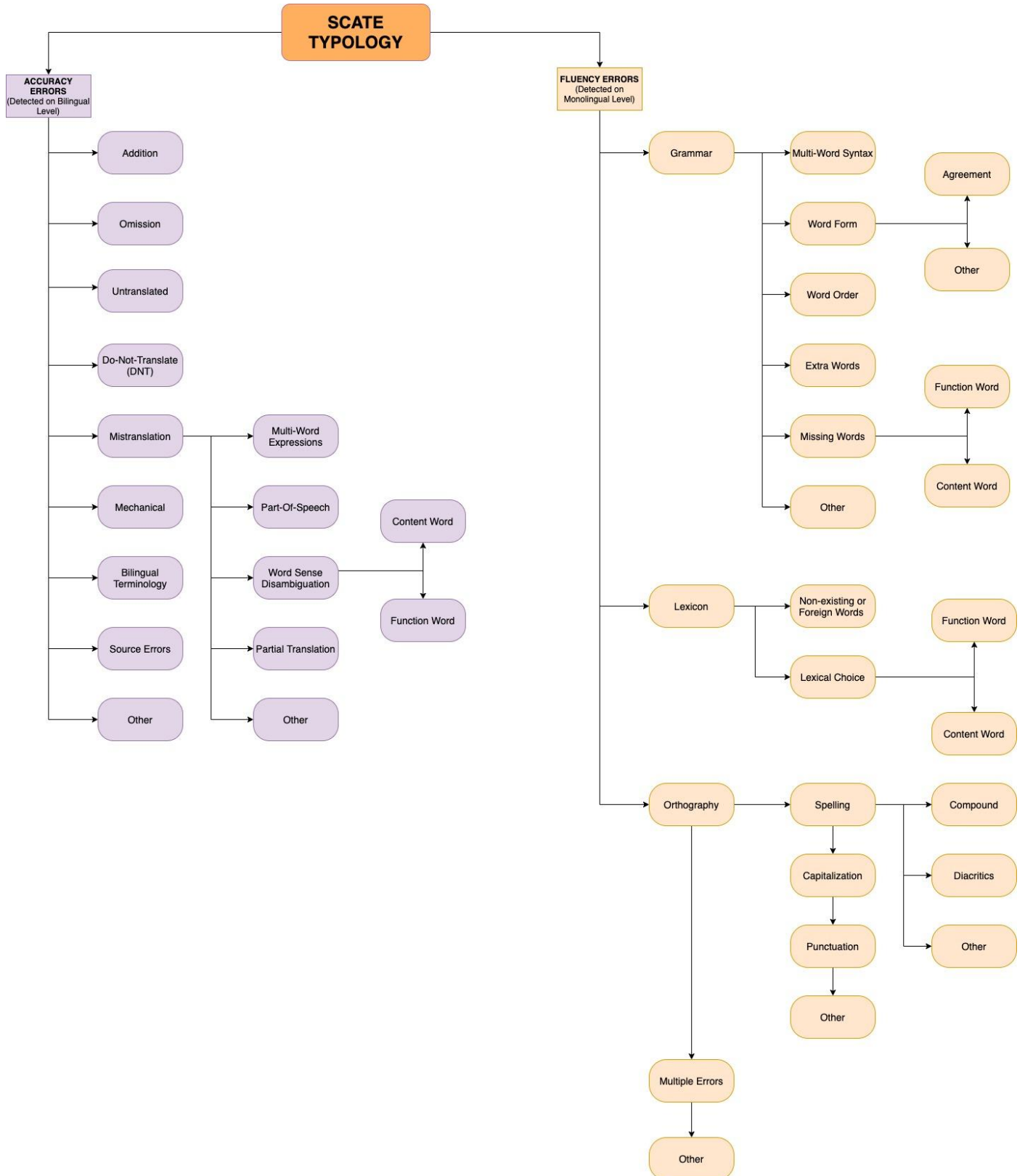
Smith, T. C., & Witten, I. H. (1993, January). Language inference from function words.

<https://core.ac.uk/download/pdf/44290163.pdf>

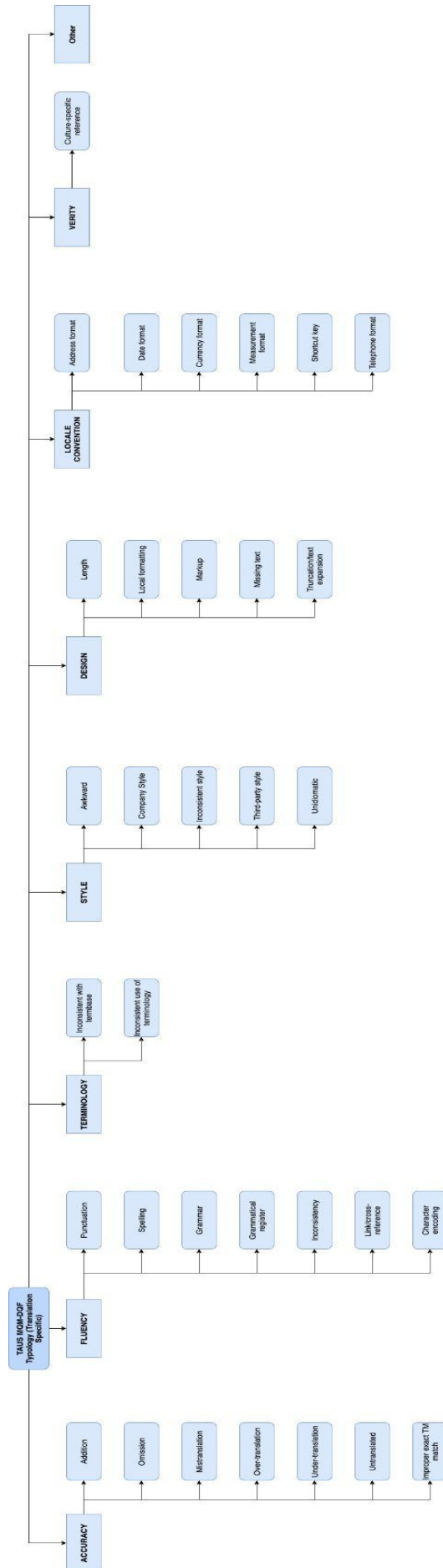
- Soloviev, K. (2017, June 27). *Measuring Content Quality with Error Typology: Step by Step Guide*. TAUS Blog - TAUS. Retrieved November, 2020, from <https://blog.taus.net/measuring-content-quality-with-error-typology-step-by-step-guide>
- Spector, N. (2017, July 23). *What the Heck? Why Are We Cursing at Customer Service Agents?* NBC News. Retrieved June 2, 2021, from <https://www.nbcnews.com/business/consumer/what-heck-why-are-we-cursing-customer-service-agents-n784836>
- Stymne, S., Pettersson, E., Megyesi, B., & Palmér, A. (2017). Annotating Errors in Student Texts: First Experiences and Experiments. *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa*(Linköping Electronic Conference Proceedings), 134: 47–60. <https://aclanthology.org/W17-0306.pdf>
- Subandowo, D. (2017, January). The Language Interference in English Speaking Skill for EFL Learners. *110*(Advances in Social Science, Education and Humanities Research (ASSEHR)), 204-208. 10.2991/iselt-17.2017.36
- Sugden, D. (1985, December). Machine Aids to Translation: Automated Language Processing System (ALPS). *Meta*. <https://www.erudit.org/en/journals/meta/1985-v30-n4-meta310/004310ar/>
- TAUS. (2015). *Harmonized DQF-MQM Error Typology*. TAUS - The Language Data Network. Retrieved November, 2020, from <https://www.taus.net/qt21-project#harmonized-error-typology>
- TeachingEnglish. *Accuracy*. TeachingEnglish British Council BBC. Retrieved 2021, from https://www.teachingenglish.org.uk/article/accuracy#skip_to_here

- Tezcan, A., Macken, L., & Hoste, V. (2017). SCATE taxonomy and corpus of machine translation errors. *Trends in E-Tools and Resources for Translators and Interpreters*. 10.1163/9789004351790_012
- Unbabel. *Seamless Multilingual Translation Services - Unbabel*. Retrieved September, 2020, from <https://unbabel.com/>
- Vasek, A. (1986, August 18-23). LINGUISTIC INTERFERENCE IN COMMUNICATION. 63-81. https://digilib.phil.muni.cz/bitstream/handle/11222.digilib/101486/A_Linguistica_39-1991-1_9.pdf?sequence=1
- Walker, P. R. (2014). *The Trials and Triumphs of Leon Dostert '28*. Wayback Machine. Retrieved January, 2021, from <https://web.archive.org/web/20160504222002/http://www.oxy.edu/magazine/fall-2015/trials-triumphs-leon-dostert-28>
- Whitlam, J. (2011). *Modern Brazilian Portuguese Grammar: A Practical Guide*. Routledge. https://nelic.ufsc.br/files/2019/04/John-Whitlam-Modern-Brazilian-Portuguese-Grammar_-A-Practical-Guide-Modern-Grammars-2010-Routledge.pdf
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. https://openreview.net/attachment?id=SkeHuCVFDr&name=original_pdf

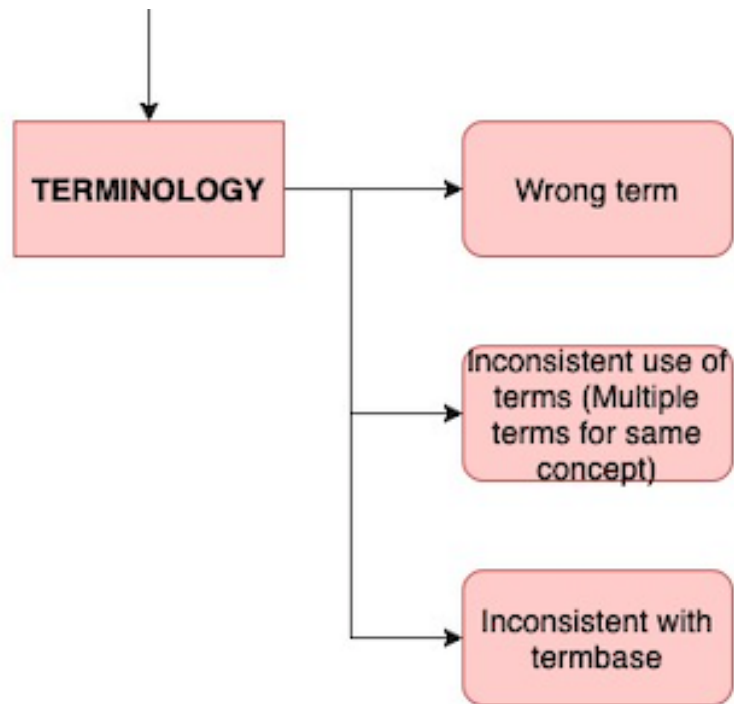
A. The SCATE MT Error Taxonomy



B. TAUS MQM-DQF Typology (Translation Specific)



C. MQM Top Level typology

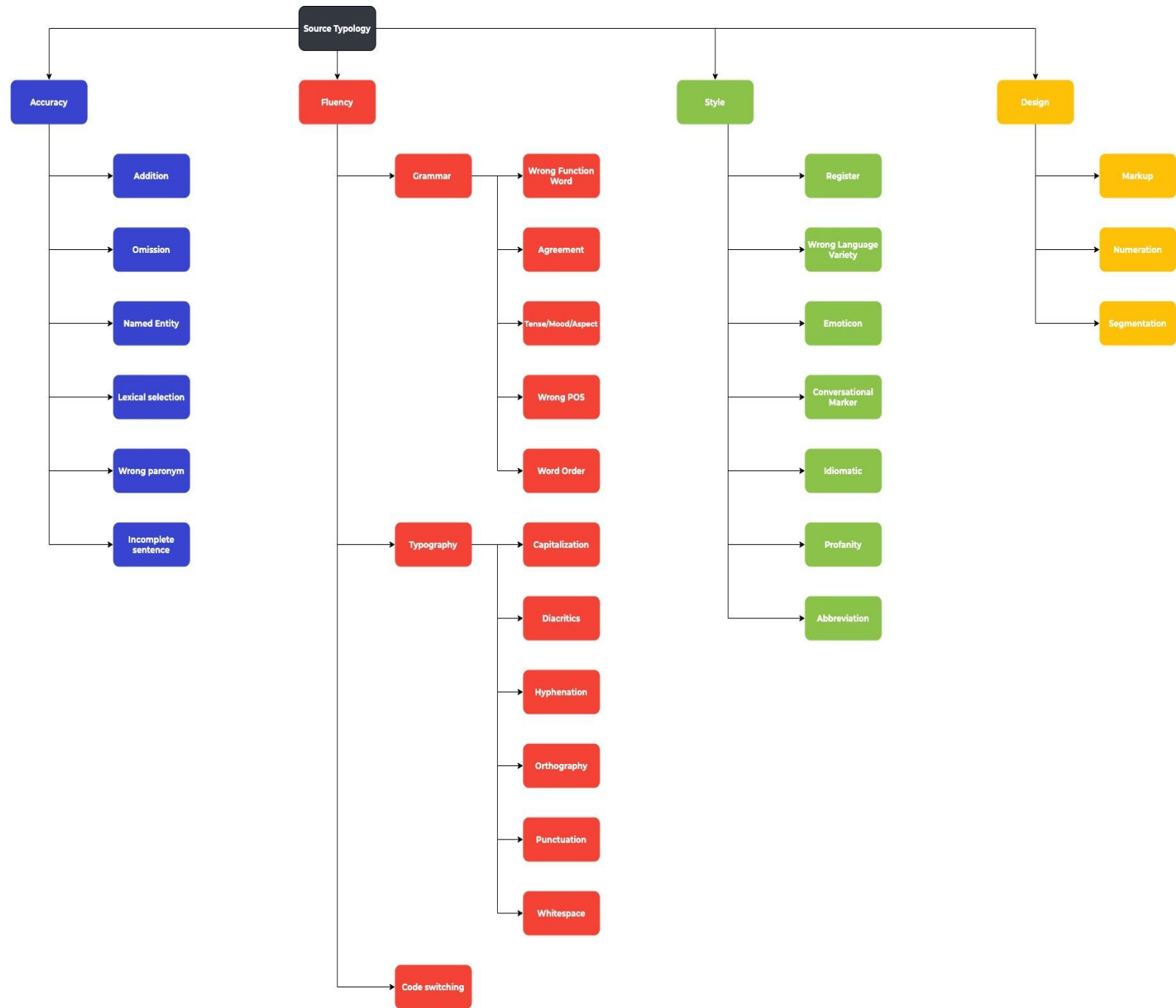


D. Typology of editing operations (Miyata & Fujita, 2021)

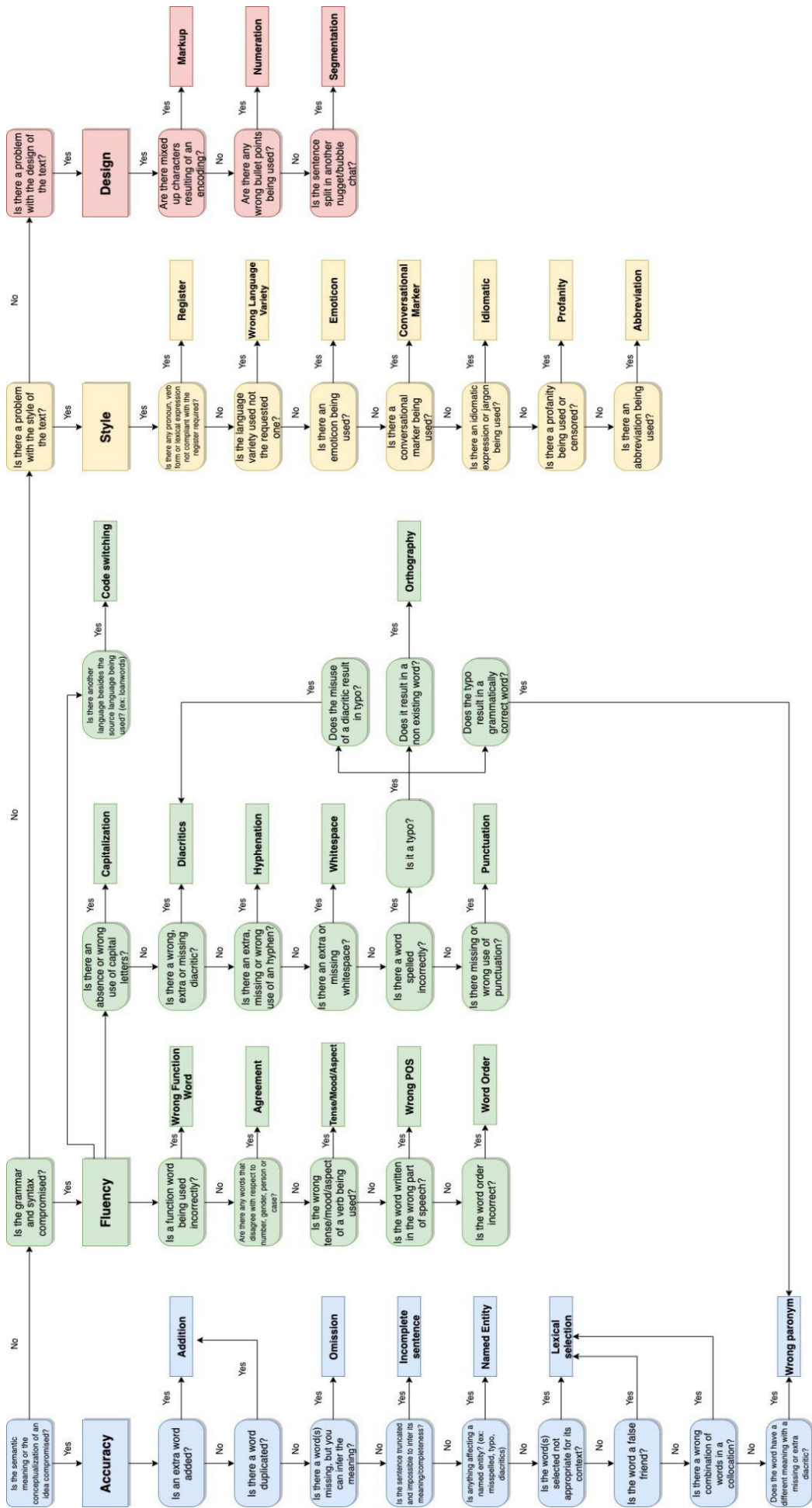
ID	Editing operation type	Ja-En		Ja-Zh		Ja-Ko		Total	Expl.	Impl.	Pres.
		G	T	G	T	G	T				
S01	Sentence splitting	1	0	3	3	4	3	14	0	0	14
S02	Structural change	3	5	9	4	4	2	27	8	1	18
S03	Use/disuse of topicalisation	1	7	4	3	1	3	19	5	2	12
S04	Insertion of subject/object	2	1	1	3	5	2	14	14	0	0
S05	Use/disuse of clause-ending noun	3	2	2	2	2	1	12	12	0	0
S06	Change of voice	1	3	0	0	0	0	4	2	0	2
S07	Other structural changes	1	0	2	1	1	0	5	3	0	2
P01	Insertion/deletion of punctuation	19	16	5	12	9	10	71	0	0	71
P02	Use/disuse of chunking marker(s)	6	12	2	1	3	4	28	11	8	9
P03	Phrase reordering	6	4	7	1	9	4	31	0	0	31
P04	Change of modification	1	3	3	0	0	0	7	0	0	7
P05	Change of connective expression	3	18	4	2	10	3	40	24	5	11
P06	Change of parallel expression	3	8	2	8	4	11	36	7	2	27
P07	Change of apposition expression	1	7	2	1	1	4	16	8	4	4
P08	Change of noun/verb phrase	1	3	2	1	3	3	13	9	3	1
P09	Use/disuse of compound noun	1	5	2	2	6	12	28	16	12	0
P10	Use/disuse of affix	4	4	1	2	3	3	17	1	0	16
P11	Change of sahen noun expression	0	1	1	1	2	0	5	1	0	4
P12	Change of formal noun expression	1	2	2	2	2	0	9	4	0	5
P13	Other phrasal changes	0	1	0	1	2	1	5	4	0	1
C01	Use of synonymous words	18	18	19	18	25	20	118	14	10	94
C02	Use/disuse of abbreviation	2	7	2	2	1	7	21	19	2	0
C03	Use/disuse of anaphoric expression	4	4	2	2	1	1	14	10	2	2
C04	Use/disuse of emphatic expression	1	2	2	1	4	1	11	10	1	0
C05	Category indication/suppression	5	3	6	5	4	7	30	29	1	0
C06	Explanatory paraphrase	3	4	1	0	1	1	10	0	0	10
C07	Change of content	22	20	21	9	14	8	94	57	23	14
F01	Change of particle	9	14	4	6	7	7	47	13	5	29
F02	Change of compound particle	8	5	5	2	5	6	31	24	2	5
F03	Change of aspect	1	4	1	0	5	1	12	0	0	12
F04	Change of tense	0	0	1	1	1	1	4	0	0	4
F05	Change of modality	3	1	2	1	3	1	11	5	0	6
F06	Use/disuse of honorific expression	3	1	1	2	2	1	10	0	0	10
O01	Japanese orthographical change	10	16	9	5	9	12	61	12	4	45
O02	Change of half-/full-width character	0	5	3	2	2	4	16	7	1	8
O03	Insertion/deletion/change of symbol	0	2	0	0	0	0	2	0	0	2
O04	Other orthographical change	0	1	0	0	3	0	4	0	0	4
E01	Grammatical errors	0	8	5	2	2	5	22	-	-	-
E02	Content errors	5	0	8	1	1	1	16	-	-	-

Table 5: Constructed typology of editing operations (G: Google, T: TexTra). The first letter of ID indicates the six major categories (S: Structure, P: Phrase, C: Content word, F: Functional word, O: Orthography, E: Errors casually introduced in the ST). The right three columns provide the frequencies for general informational strategies (Expl.: Explicitation, Impl.: Implication, Pres.: Preservation).

E. Source Typology



F. Source Typology decision tree



G. Severities' decision tree

