Lessons Learned: Problems, Challenges and Learnings of an Annotator Training Initiative

Érica Sofia Sampaio de Lemos

Relatório de Estágio de Mestrado em Tradução

Especialização em Inglês

May 2023

Relatório de Estágio apresentado para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Tradução realizado sob a orientação científica do Prof. Doutor Marco Neves e da Doutora Marina Sánchez Torrón.

Acknowledgments

Ao meu orientador, Professor Doutor Marco Neves, pelas mil e uma reuniões a esclarecer dúvidas e a resolver problemas e por toda a ajuda.

À Marina, por toda a paciência e apoio. Sem si este estágio e este relatório não teriam sido possíveis.

À Unbabel, especificamente à *Community Team* por me terem recebido tão bem e por estarem sempre prontos a ajudar.

Ao grupo de TTT por todos os risos e por me lembrarem de que há duas formas de fazer as coisas e que, infelizmente, o Raul é que está certo. Um agradecimento especial à Sofia, que me acompanhou durante esta experiência toda e esteve sempre pronta para stressar comigo.

À Marta, por estar sempre pronta para me ouvir e por me oferecer sempre um espaço em que posso relaxar.

À Inês, ao António e à Raquel, que apesar da distância geográfica ou mental (sim, António), são sempre um grande apoio.

Ao Filipe, por me fazer sempre sentir melhor quando não sei o que estou a fazer e duvido de mim mesma.

Aos meus pais, por terem paciência e me ouvirem naqueles dias em que parece que tudo corre mal e em que o meu stress não vos permite chamarem-me à razão.

LESSONS LEARNED:

PROBLEMS, CHALLENGES AND LEARNINGS OF AN ANNOTATOR TRAINING INITIATIVE

Érica Sofia Sampaio de Lemos

Abstract

Evaluating the quality of translations presents some challenges. Manual processes are said to be the most comprehensive but are not free of problems, like evaluator subjectivity or lack of specialized knowledge. The goal of this report is to explore an annotator training initiative to improve reliability of quality annotations at Unbabel. Due to the introduction of a new error typology in March 2022, Unbabel had found that some annotators were missanotating. In an attempt to address this issue, a hypothesis arose: by creating an annotation training focused on problem areas, the reliability of annotations will improve.

To test this hypothesis, we conducted a data-oriented investigation and by analyzing various annotation assignments and relying on specialist internal knowledge it was possible to identify the error types that were causing the most confusion. With the data gathered, a question-based annotation training was developed with the objective of assessing the knowledge annotators had of Unbabel's annotation guidelines. This training was sent to annotators who were found to be missanotating on a regular basis and, after the completion of the training, an observational study was held in which annotation assignments of six annotators were analyzed before and after taking the training in order to gauge the training's potential effectiveness.

Although the results showed that one-off training was insufficient to improve annotation reliability, the work developed did generate ideas on how to make future trainings more effective, as well as ideas on how to incorporate the learning process into actual annotation assignments by making changes to the user interface and, as a result, creating a more integrated learning experience.

KEYWORDS: Machine Translation; Annotation; Annotation Training; Observational Study; Quality Processes.

LESSONS LEARNED:

PROBLEMS, CHALLENGES AND LEARNINGS OF AN ANNOTATOR TRAINING INITIATIVE

Érica Sofia Sampaio de Lemos

Resumo

A avaliação da qualidade de traduções apresenta certos desafios. Os processos manuais são considerados os mais abrangentes, mas não estão isentos de problemas, como a subjetividade do avaliador ou a falta de conhecimentos especializados. O objetivo desta tese é explorar uma iniciativa de formação para anotadores com a finalidade de melhorar a fiabilidade da qualidade das anotações na Unbabel. Devido à introdução de uma nova tipologia de erros em março de 2022, a Unbabel descobriu que alguns anotadores anotavam incorretamente. Numa tentativa de resolver este problema, surgiu uma hipótese: ao criar uma formação para anotadores centrada em áreas problemáticas, a fiabilidade das anotações melhorará.

Para testar esta hipótese, realizámos uma investigação orientada para os dados e, ao analisar várias tarefas de anotação e basearmo-nos em conhecimentos internos especializados, foi possível identificar os tipos de erro que causavam mais confusão. Posteriormente à recolha de dados, foi desenvolvida uma formação para anotadores que consiste em perguntas com o objetivo de avaliar o conhecimento que os anotadores tinham das diretrizes de anotação da Unbabel. Esta formação foi enviada a anotadores que anotavam incorretamente com regularidade e, após a conclusão da formação, foi realizado um estudo observacional, no qual foram analisadas as tarefas de anotação de seis anotadores antes e depois de completarem a formação, com o objetivo de avaliar a potencial eficácia da formação.

Embora os resultados tenham demonstrado que uma formação pontual não foi suficiente para melhorar a fiabilidade das anotações, o trabalho desenvolvido gerou ideias sobre como tornar futuras formações mais eficazes, bem como ideias sobre como incorporar o

processo de aprendizagem em tarefas de anotação reais, fazendo alterações à interface do utilizador e, consequentemente, criando uma experiência de aprendizagem mais integrada.

PALAVRAS-CHAVE: Tradução Automática; Anotação; Formação para anotadores; Estudo Observacional; Processos de Qualidade.

List of acronyms

AI - Artificial Intelligence

MT - Machine Translation

FAQ - Frequently Asked Questions

GDPR - General Data Protection Regulations

NER - Named Entity Recognition

QE - Quality Estimation

MQM - Multidimensional Quality Metrics

TQE - Translation Quality Evaluation

CUA - Customer Utility Analysis

ALPAC - Automatic Language Processing Advisory Committee

TAUM - Traduction Automatique de l'Université de Montréal

CEC - Commission of the European Communities

RBMT - Rule-based machine translation

ST - Source Text

TL - Target Language

SMT - Statistical Machine Translation

NMT - Neural Machine Translation

LLMs - Large Language Models

TQA - Translation Quality Assessment

LSP - Language Service Provider

LISA - Localization Industry Standards Association

TAUS - Translation Automation User Society

DQF - Dynamic Quality Framework

DFKI - German Research Centre for Artificial Intelligence

LP - Language Pair

BLEU - Bilingual Evaluation Understudy

METEOR - Metric for Evaluation of Translation with Explicit Ordering

COMET - Crosslingual Optimized Metric for Evaluation of Translation

UI - User Interface

CONTENTS

1. INTRODUCTION

This report was performed in the context of the Master's degree on Translation of NOVA FCSH. To this end, an internship was carried out at Unbabel—a software company that provides translation services in the area of customer support—, due to the company's innovative strides in the world of translation. This internship was completed concurrently with my colleague Ana Sofia Martins, who also interned at Unbabel.

Unbabel's translation services are provided by using a mixture of human quality processes—among these, error annotation—and automatic quality processes. The work developed throughout this internship was in the area of error annotation, a process that consists of identifying translation errors and using a specific annotation error typology to categorize them. The error typology that is currently in use at Unbabel is an MQM-based typology. In 2022, after Unbabel acquired Lingo24—a tech-enabled translation and localization company that worked in areas outside of customer support—, a new typology was devised to accommodate the new content types that came with Lingo24. However, this change in the typology created a problem with the reliability of annotations since annotators needed time to get accustomed with this new typology.

The scope of this report is to improve the reliability of annotations. For this reason, a question-based annotation training was developed and sent to annotators who were found to be missanotating regularly, with the objective of providing annotators with the necessary knowledge and skills to reduce errors and enhance the quality of annotations.

After designing the training and having a set of annotators completing it, an observational study was held by analyzing annotation assignments from before and after annotators took the training, to gauge the training's potential effectiveness. For this study, we chose six annotators who worked with language pairs for which there were internal language experts: EN-PT-BR, EN-PT, EN-DE and EN-NL.

In Section 2 of this report, there is some background on Unbabel, focusing on the content types that the company translates and the quality processes that are at work to provide these translation services. Section 3 then begins with an overview of Machine Translation's history as well as some of the human and automatic quality methods that have been created during the course of the development of Machine Translation. Section 4 briefly underlines the

objectives of the internship that was held at Unbabel and in Section 5 the methodology for the work done in this internship is explained, going over the creation of an annotation training and an FAQ, and the development of an observational study. In Section 6, there is a detailed analysis of the results from the observational study and, finally, Section 7 highlights the conclusions and limitations found, as well as a prospect for future work.

2. HOST CHARACTERIZATION – UNBABEL

Unbabel is a Portuguese software company that was founded in 2013 by Vasco Pedro, João Graça, Hugo Silva, Sofia Pessanha, and Bruno Prezado. Its main focus is on providing translation services, with a particular emphasis on customer support, while combining both artificial intelligence (AI) and human translation to deliver fast and accurate translations.

While a big worry for translators nowadays is if AI will replace human translation due to its speed and inexpensiveness, Machine Translation (MT) is still far from perfect and can, at times, be unreliable. Which is why Unbabel chooses to combine both, working with professional translators and qualified linguists to create a community of freelance post-editors and annotators who review machine translation. The company's unique approach to translation enables the connection between companies and their customers by eliminating communication barriers, while delivering faster and high-quality translations.

Unbabel had been offering services in 27 language pairs and had offices in San Francisco, New York, and hubs in London and Berlin, in addition to its headquarters in Lisbon. However, in December of 2021, Unbabel acquired Lingo24, a tech-enabled translation and localization company based in Edinburgh, with offices in Cebu (Philippines) and Timișoara (Romania). The acquisition broadened the scope of possibilities for Unbabel, due to Lingo24's history of offering translation services in fields other than customer support in various language pairs, while working with a big community of translators, transcreators, and copywriters. After this acquisition, Unbabel now works with over 60 language pairs.

The sections that follow in this chapter will explain Unbabel's translation and quality processes. Section 2.1. will discuss the translation workflows used at Unbabel. Following that, Section 2.2. will outline Unbabel's quality processes.

2.1. Translation Workflows

Before the integration of Lingo24, the content types translated by Unbabel revolved around the area of Customer Support, namely: tickets, chat and Frequently Asked Questions (FAQs). Unbabel's translation workflows are represented through pipelines, each content type

mentioned being represented through a different pipeline. The following sections were based on the reports written by Beatriz Silva, Madalena Gonçalves and Marjolene Paulo[1].

2.1.1. Ticket translation pipeline

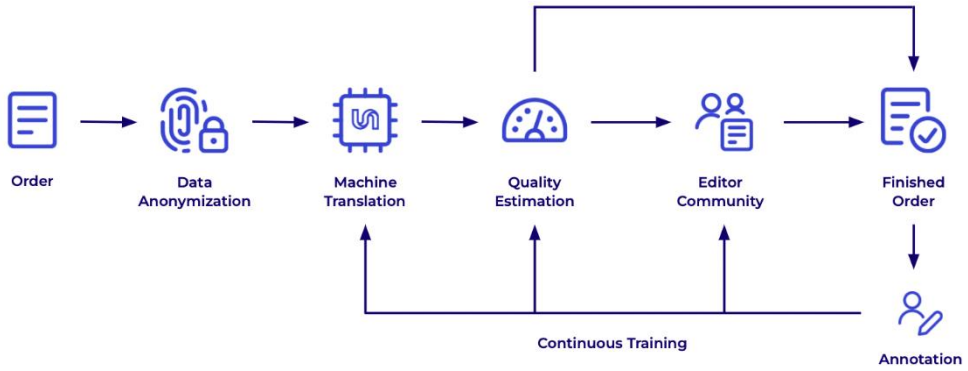


Figure 1 - Unbabel's Translation pipeline for tickets (internal resources)

Figure 1 depicts the Unbabel ticket pipeline, which includes a critical step after the client submits their order: data anonymization. When dealing with the area of Customer Service there can be access to sensitive information and, as such, it is vital to safeguard the client's privacy by following the General Data Protection Regulations (GDPR). All sensitive information, such as names, addresses, phone numbers, and passwords, is identified and replaced using Unbabel's Named Entity Recognition (NER) system. Only then is an anonymized version of the text machine translated.

Once the text is translated, there is a quality estimation (QE) step where the quality of the translation is automatically detected, and a score is attributed. If the score attributed is above a certain threshold, the translation is sent directly to the client, an action known as a QE skip, since the other steps are bypassed. However, if the QE score is lower, the translation gets sent to post-editors, who correct any errors found in the machine translated text before it is sent to the client.

The annotation process is the final step in the pipeline. This step consists of the text being sent to the community of annotators that will label the errors found in the version of the text that was sent to the client. In addition to the final translation being annotated, MT content goes through the annotation process as well, this is done to keep track of the quality delivered

[1] See Silva (2022), Gonçalves (2021) and Paulo (2022).

to editors. So not only does the process of annotation aid the editor community, but it also contributes to bettering the QE system and, most importantly, the machine translation system.

2.1.2. Chat translation pipeline

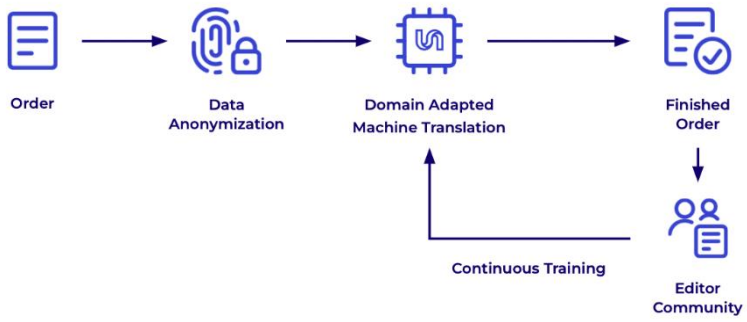


Figure 2 - Unbabel's Translation pipeline for chat (internal resources)

Figure 2 portrays the pipeline for chat, which varies from the tickets pipeline. Although it starts with the same data anonymization step, since chat is a form of communication that takes place in real time, the priority with this content type is the rate of delivery, as opposed to the quality of the translation. Seeing as there is a need for a rapid response, the translation for chat is based only on domain-adapted MT models, meaning that QE and editors are not used. Instead, after the MT, the translation is immediately sent to the client and is then sent to annotators, so the errors can be evaluated and then fed to train the MT models for improvement.

2.1.3. FAQs translation pipeline

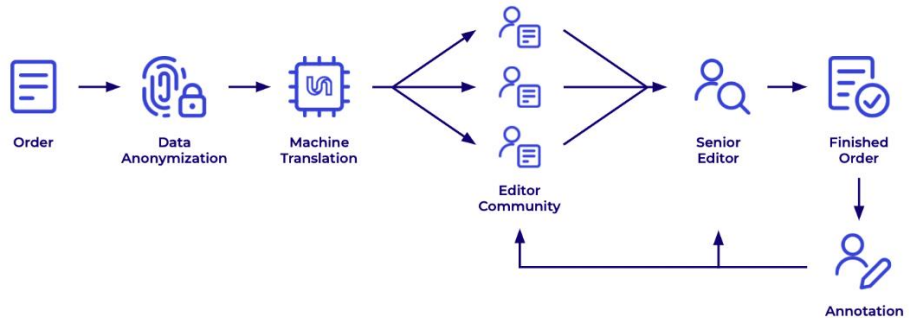


Figure 3 - Unbabel's Translation pipeline for FAQs (internal resources)

The last content type is the most quality-oriented one, seeing as a lot of times it will be posted on the client's website, becoming a part of their image and it will have a much broader audience than other content types. For this reason, after the text is anonymized and translated by the MT system, it is split into sections and the sections are sent to editors individually for post-edition—as is demonstrated in Figure 3. Once the text is reviewed and edited, the sections are reassembled, and the full document is sent to a senior editor so that it can be proofread. Only after all these steps, the document is sent to the customer.

2.2. Quality processes at Unbabel

There are various quality processes at work at Unbabel. The company works with a community of post-editors, senior editors, terminologists, evaluators, and annotators that help ensure and control for quality. However, Unbabel also uses multiple automatic quality processes for the same quality assurance and control purposes.

The following sections focus on Unbabel's quality processes, Section 2.2.1. going into detail on the annotation process, touching lightly on Multidimensional Quality Metrics (MQM) and Unbabel's annotation guidelines, as well as an analysis framework developed by Unbabel with the purpose of giving clients a visual illustration of the quality of the translation they are receiving. In Section 2.2.2., there is an explanation of the training and evaluations post-editors go through from the point of their onboarding. Lastly, in Section 2.2.3., the automatic quality processes that have been developed by Unbabel—OpenKiwi and COMET —are explained.

2.2.1. Annotation process

The process of identifying translation errors and categorizing them according to a specific annotation error typology is known as error annotation. Unbabel uses an adaptation of the MQM[2] framework to classify errors, assign them severities and based on those, obtain a numerical score that represents quality. According to its official definition, MQM is "a framework for analytic Translation Quality Evaluation (TQE)"[3], meaning that it is used to evaluate the quality of a translation. The MQM framework will be discussed in more detail in

[2] Check "Lommel, A., et al. (2014). Multidimensional quality metrics (mqm): a framework for declaring and describing translation quality metrics. *Revista Tradumàtica: Tecnologies de La Traducció*, (12), 455–463. https://doi.org/10.5565/rev/tradumatica.77"
[3] https://themqm.org/

Section 3.2.1., however, it is important to mention that this framework was designed with the intention of being adaptable, each user being able to choose the error types that best accommodate the content types they will be working with and the severities that can be attributed to each error. The typology in use at Unbabel is comprised of eight parent categories: Accuracy, Linguistic Conventions, Terminology, Design and Markup, Style, Locale Conventions, Audience Appropriateness and Custom; and 31 children categories, as is shown in Figure 4[4].

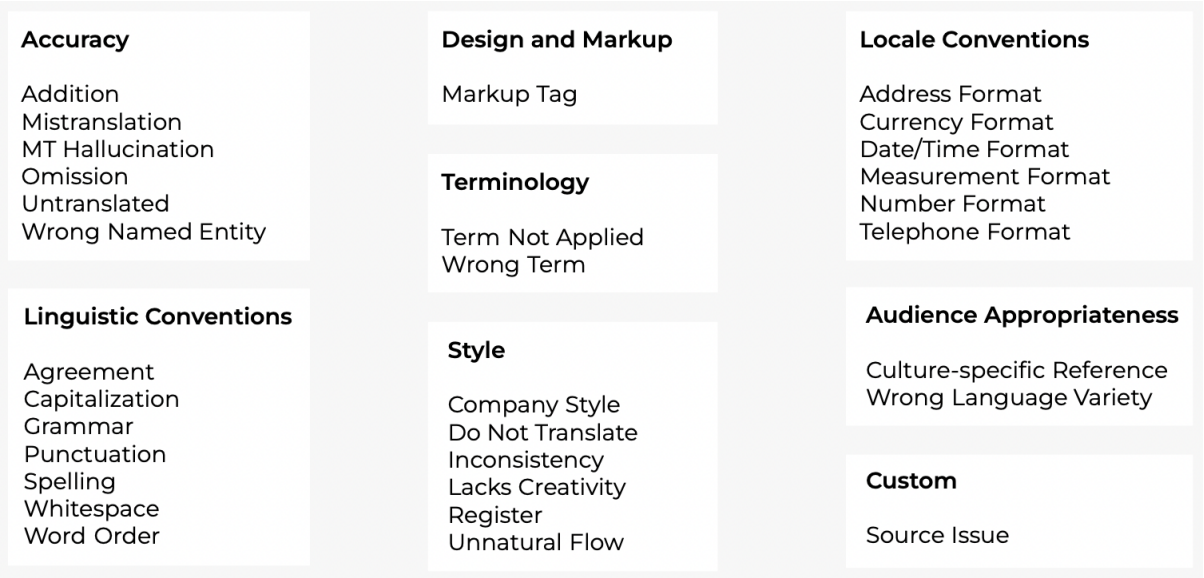


Figure 4 - Unbabel's error typology

Furthermore, a severity is assigned to each error depending on how much it impacts the translation. The four severities are: Neutral, Minor, Major and Critical. The Neutral severity is, at the moment, reserved only for the Source Issue category, and it is used when there is an error in the Source Text. A Source issue annotated with the Neutral severity will not affect the translation, but it will help to determine which errors in the target text are caused by issues in the source.

If an error does not impact the actual content of the text and only makes it less fluid, the Minor severity is attributed. Examples of Minor severity errors include, often, Punctuation errors. Major errors may cause some meaning to be lost but do not render a translation unusable.

[4] Unbabel's annotation guidelines are protected by an NDA so it's not possible to disclose them here. However, themqm.org contains definitions for most of these categories that are consistent with Unbabel's definitions.

A Major error would be, for example, a Grammar error that makes the text difficult to understand. The Critical severity is attributed when the error severely changes the meaning of the original text, misleading the reader or even leading to security or legal complications, or when it appears in a prominent part of the text. A spelling error in a press release headline can be assigned a Critical severity.

Each severity has a corresponding penalty multiplier so that the MQM score can be calculated, with Critical being attributed 10 points, Major 5 points and Minor 1 point. The MQM formula takes into account the number of words. To calculate the MQM score of a 250-word document with 2 major errors and 3 critical error, the formula would be:

=100-(100*(3*10 + 2*5)/250) = **84 MQM**

At Unbabel, all annotations are done on an in-house platform called the Annotation Tool. The annotation process is a very important part of Unbabel's pipeline, because not only does it bring an insight into the quality and performance of the MT by helping identify error patterns, but, since not only MT but also the post-edited text is annotated, it also allows the Community Team[5] to filter out editors who are not delivering quality translations. Annotation also allows Unbabel to develop a dynamic quality framework for measuring and communicating quality to clients based on the content type and quality level agreed upon with the client.

In order to easily communicate to clients, the quality of the translations they receive, Unbabel developed an analysis framework for quantifying and visualizing translation quality, since raw MQM scores had little value for customers, and they reported difficulties interpreting them. This analysis framework is called Customer Utility Analysis (CUA) and aggregates MQM scores into five different quality buckets: Best, Excellent, Good, Moderate and Weak. Each bucket is determined by a range of MQM scores, as can be seen in Figure 4.

[5] The work done by the Community Team—in which the internship took place—will be explained in Section 2.3.

Best	**Premium quality (MQM of >98)** The translation is of optimal quality. Minor errors are possible but they are inconsequential and extremely sparse.
Excellent	**High level of quality (MQM in the range 94 ≤ x < 98)** The translation sounds fluent. Only a few minor style or fluency issues are still to be expected but the meaning is preserved and communication is successful.
Good	**Good level of quality (MQM in the range 80 ≤ x < 94)** The translation is understandable. There might be grammatical issues or occasionally inaccuracies in meaning.
Moderate	**Below target quality, area of attention (MQM in the range 60 ≤ x < 79)** The understandability of the translations in this zone might be disrupted due to the nature and the amount of the errors.
Weak	**Poor quality (MQM of <60)** The translation contains errors which likely impede understanding of the core concepts of the message.

Figure 5 – Unbabel's Customer Utility Analysis framework (internal resources)

2.2.2. Post editors' evaluation

Unbabel works with various communities that include editors and senior editors, as well as what is called the PRO Community: terminologists, evaluators and annotators. Unbabel conducts pre-onboarding and ongoing evaluations of the post-editor community since they play a vital role in the quality of the final product that is sent to the client.

Editors start by taking a timed multiple-choice language test to assess their skills and attention to detail (they will take one language test per language pair they wish to translate), which is followed by a tour of the interface they will be using when post-editing, and after this, they have to complete a skill test. Once they have completed all of these steps, they have to complete five training tasks for each language pair they chose to work with. After these training tasks are evaluated, if their work is satisfactory, they move on to paid tasks. However, in order to ensure that the quality is preserved, editors who move on to paid tasks will still go through periodical evaluations and depending on the result, they will either continue working on paid tasks or they will be demoted back to training tasks. These evaluations are carried out by randomly selecting a task done by a post-editor and assigning it to an Evaluator[6], who will then represent the quality of these tasks through a star rating system from 1 (a bad quality translation) to 5 (an excellent translation), as shown in Figure 5.

[6] Evaluators are professional translators or linguists.

Figure 6 – Post-editors' quality rating system (internal resources)

2.2.3. Automatic Quality Metrics

Unbabel manages a large number of translations due to their many clients and, due to the fact that human evaluation can be very time consuming and oftentimes expensive, the company has several automatic metrics in place to assess the quality of the translations made by their MT systems.

As was seen in Unbabel's translation pipelines' (Section 2.1.), the step of QE measures the quality of translations and, at Unbabel, QE is done automatically through an in-house software named OpenKiwi. OpenKiwi is a "source framework for QE that implements the best QE systems from WMT 2015-18 shared tasks, making it easy to combine and modify their key components, while experimenting under the same framework" (Kepler et al., 2019). Kiwi automatically assesses translation quality and assigns a score based on whether the translation is favorable or not. If the attributed score is higher than a particular threshold, no post-edition is required, and the translation is sent to the client.

COMET is another important in-house developed metric which, as will be further discussed in Section 3.2.2., is a "neural framework for training multilingual machine translation evaluation models" (Rei et al., 2020) and is used to measure the quality of the MT systems. Unlike previous frameworks with the same goal, "COMET captures the meaning similarity between texts with enough granularity to accurately predict human experts' translation quality judgments." ("Unbabel Launches COMET, Blazing New Trail for Ultra-Accurate Machine Translation", 2022).

2.3. Community Team

The Community Team’s role focuses, as the name implies, on the communities that help Unbabel deliver quality translations: editors, senior editors, evaluators, terminologists, and annotators. The team's goal is to achieve a balance between three elements when delivering translations: speed, quality, and cost. They focus their efforts on what matters most to Unbabel when it comes to translations involving humans: quality. They also support the community in their work by creating resources and training materials and implementing quality of life changes to the platforms to make tasks easier and faster to complete.

3. STATE OF THE ART

This chapter will provide an overview of the history of machine translation, followed by an explanation of the three types of rule-based machine translation systems (Section 3.1.1.): direct, transfer and interlingual; and the three sub-types of data-ruled machine translation (Section 3.1.2.): example-based MT, statistical MT and neural MT. In Section 3.2. there will be an examination of the quality methods that have been established throughout the years in an attempt to assess the quality of MT outputs; this examination will be divided into two sections: manual quality metrics (Section 3.2.1.) and automatic quality metrics (Section 3.2.2.).

3.1. Machine Translation

Although Hutchins (2003) acknowledges seventeenth century ideas of "universal (and philosophical) languages of 'mechanical' dictionaries'", the author notes that the first true proposals of machine translation were only made in the twentieth century. In 1933, two patents were filled for "electromechanical devices" that could be considered as mechanical dictionaries, these were filled by Georges Artsrouni (France) and Petr Trojanskij (Russia) (Kenny, 2018).

However, only in 1946 were ideas put forward about using computers for purposes of translation, by Andrew Booth (a British computer scientist) and Warren Weaver (of the Rockefeller Foundation) and in 1949 a Memorandum written by Weaver is circulated, where he gives a proposal on how to conquer the "world-wide translation problem" by use of electronic computers: in order to address ambiguity issues and the limitations of word-for-word translation, he recommends using context of occurrences of a particular word as a method of disambiguation by examining the words used before and after this word (Kenny, 2018; Hutchins, 2003). Although Weaver faced some skepticism after the publishing of his Memorandum, it would serve as a starting point for MT research in the United States.

In 1951, the first MT conference was held at MIT, which discussed topics that would later become vital in the area of MT: pre-editing and post-editing (Kenny, 2018). Among the many subjects discussed at the conference, León Dostert from Georgetown University suggested that a public demonstration of the functionality of MT might be needed to attract research funding (Hutchins, 2003). Accordingly, on 7 January 1954, a demonstration of a Russian to English MT was made by Léon Dostert in collaboration with IBM and, although the

system presented its limitations, it was sufficient to encourage significant financing for MT research in the United States, some people even believing that MT systems would become “capable of translating almost everything” within five years (Kenny, 2018; Hutchins, 2003).

Following the Georgetown demonstration, numerous research groups began to appear in the United States, the Soviet Union, and other countries. Due to the Cold War, most research in the United States was supported by the CIA and the military and focused on translating Russian writings into English; while research in the Soviet Union seemed to focus on a far broader variety of language pairs (Kenny, 2018). During this time, three basic approaches to MT research emerged: the "direct translation" model, the "interlingua" model, and the "transfer approach" (Hutchins, 2003). Section 3.1.1. will go into more detail about these three different MT systems.

Even though in the 1950s researchers were confident in the prospects of MT research, the 1960s proved to be a difficult period in this regard considering "the imminent prospect of good quality MT was receding" (Hunchins, 2003). In 1960, the inability of MT to resolve semantic ambiguities led Bar-Hillel to argue that MT had reached a stalemate, claiming that MT would never have the encyclopedic knowledge of a human and that “Fully automatic, high quality translation is not a reasonable goal, not even for scientific texts… Reasonable goals are then either fully automatic, low quality translation or partly automatic, high quality translation.” (Bar-Hillel, 1960), implying that translators would play the role of post-editors to correct MT output (Kenny, 2018).

The event that truly stalled MT research progress was, however, when the Automatic Language Processing Advisory Committee (ALPAC) was asked, by the government sponsors of MT in the United States, to examine the less-than-favorable results in MT research (Kenny, 2018; Hutchins, 2003). The Committee concluded, in 1996, that “MT was slower, less accurate and twice as expensive as human translation” (Hutchins, 2003) and that “there is no immediate or predictable prospect of useful machine translation” (ALPAC, 1966), recommending instead that machine aids for translators be developed and that efforts be redirected towards research in computational linguistics (Hutchins, 2003). Despite widespread criticism that the ALPAC Report was "biased and short-sighted", its impact was substantial, causing a halt in MT research in the United States and the Soviet Union—which determined that their prospects of success were even lower due to poorer computer facilities (Hutchins, 2003).

Despite the consequences of the ALPAC Report, growing interest in the area of MT started showing in Canada, after the Official Language Act in 1969 that officialised the bilingualism[7] in the country—creating a demand for English-French translations (Kenny, 2018). The TAUM (Traduction Automatique de l'Université de Montréal) project developed the Météo system, which is recognized as one of the twentieth century's most effective MT implementations, with an English-to-French version in service from 1977 to 2002 and translating up to 45,000 words per day (Kenny, 2018; Hutchins, 2003; Langlais et al., 2005).

The European Union, in turn, began using MT in 1976, when the Commission of the European Communities (CEC) began using the Systran MT system for English-French translations (which were then followed by systems for translating most other European Union languages) (Hutchins, 2003). Peter Toma created Systran's system, with its oldest version, the Russian-English system, being utilized by the other organizations in the United States (Kenny, 2018; Hutchins, 2003). The European Union continued to utilize Systran until 2010, up until the European Commission created mt@ec, an in-house MT system, which was eventually replaced in 2017 by another in-house system named eTranslation (Kenny, 2018).

With the swift adoption of personal computers and the development of other technological breakthroughs, such as integrated tools at the translator's workstation, the 1980s and 1990s saw the emergence of various commercial MT systems (Hutchins, 2003).

While the goal of MT development in its initial stages was to achieve totally automatic quality translations (Hutchins, 2003), the current era has brought technological improvements and increased usage of the internet, opening up new opportunities for MT. More complex systems that provide higher quality translations at a faster rate have emerged since the 2000s. In Section 3.1.1., the previously mentioned systems—direct systems, transfer systems, and interlingual systems—are explained in more detail. In Section 3.1.2., three of the more complex systems that gained traction in the 2000s will be discussed: example-based MT, statistical MT and neural MT.

3.1.1. Rule-based machine translation

As has been explained throughout this chapter, before the turn of the century, MT has relied on an approach that "manipulates linguistic knowledge in the form of handcrafted

[7] English and French became the official languages of Canada.

grammatical and lexical rules" (Kenny, 2018). The systems that have been presented are commonly known as "rule-based machine translation" systems (or RBMT systems) and were state-of-the-art up until the 2000's. RBMT systems can be divided into three different types: direct, transfer and interlingual (Hutchins, 2003).

The direct systems are indicative of what was deemed to be the first generation of MT, with examples comprising pre-ALPAC systems (Kenny, 2018). These were multilingual systems "with a minimal amount of analysis and syntactic reorganisation" (Hutchins, 2003), consisting of word-for-word translations with no regard for context (Kenny, 2018). These systems were quite limiting, as they only used a few lexical rules and allowed only for some local reordering of words, which was insufficient to eliminate errors of ambiguity (Kenny, 2018).

To address the shortcomings of the direct approach, transfer systems were developed during the second generation of MT, an approach which consisted of three stages: analysis, transfer, and generation (Hutchins, 2003). The transfer approach started with the premise that higher-quality translations could be achieved if the source text (ST) was first syntactically analysed, with semantic issues dealt with subsequently (Kenny, 2018; Hutchins, 2003). After the analysis step was done, the ST would be converted into its syntactic structure, which would then be converted into the equivalent syntactic structure in the target language (TL) (Kenny, 2018); this would be called the transfer stage. Lastly, the generation stage, where this structure would be translated into the TL by the systems (Kenny, 2018). However, while first generation systems were lax in rules, the transfer approach could become complex due to the unlimited set of rules, that would sometimes contradict each other and pose new problems (Stein, 2018).

The last approach: the interlingual approach, started out with a similar analysis to the one used in the transfer systems, however, this analysis would "culminate in a representation of the content of the source-language sentence that no longer bore traces of the source language" (Kenny, 2018), meaning that they would use an interlingua—an artificial language devised for MT—that bore no meaning to a specific language (Koehn, 2010). Despite the efforts made to create an interlingua, it was found impractical.

3.1.2. Data-driven machine translation

Data-driven systems, also known as corpus-based systems, gained prominence around the turn of the century, eventually replacing RBMT systems. According to Kenny (2018),

"translation knowledge can be learned directly from parallel corpora (or 'bitexts'), that is, collections of source texts aligned with their human translations.", so naturally the increased usage of the internet brought forth these systems, due to the growing amount of content in different languages, which could serve as parallel corpora. Much like RBMT, data-driven MT can be divided into three sub-types: example-based MT, statistical MT and neural MT (Kenny, 2018).

The first data-drive systems: example-based systems, emerged during the 1980's and were developed essentially in Japan (Koehn, 2010). The goal of this approach was "try to find a sentence similar to the input sentence in a parallel corpus and make the appropriate changes to its stored translation" (Koehn, 2010). This approach can be compared to translation memories, which are now widely used.

The second approach, statistical MT (SMT), surfaced in the 1990s and was considered state-of-the-art until the mid-2000s, when it was replaced by neutral MT (Koehn, 2010; Kenny, 2018). Training, tuning, and decoding are the three main steps of these systems (Kenny, 2018). As data-driven systems, SMT systems learn from parallel corpora and employ a probabilistic model (or several) learnt from the parallel corpora to determine the most likely translation for a particular source sentence; this is known as the training step (Kenny, 2018). In the second step, known as tuning, "system developers work out the optimal weight that should be assigned to each model to get the best outcome" (Kenny, 2018). When it is time to translate, the system "generates many thousands of hypothetical translations for the input string and calculates which one is most probable, given the particular source sentence, the models it has learned and the weights assigned to them" (Kenny, 2018) in the decoding step.

The last approach, neural MT (NMT) is the current state-of-the-art. NMT systems have a similar learning step to the one explained for SMT, however, NMT systems use artificial neural networks that resemble human neurons[8] in the sense that "their output or activation (that is, the degree to which they are excited or inhibited) depends on the stimuli they receive from other neurons and the strength of the connections along which these stimuli are passed" (Forcada, 2017), meaning that the output delivered by these systems is predicted depending on the input they receive.

[8] For more information check Koehn (2010).

More recently, Large Language Models (LLMs) have shown state-of-the-art capabilities in text generation. The potential of LLMs to do MT is still being the subject of research efforts.

3.2. Quality processes

The previous section (3.1.) discussed the history of the progression of MT. Throughout the evolution of MT, quality assessment continually served as a decisive element in whether or not an MT system was fit for use. As a result, over the years, there has been a consistent effort to design translation quality assessment (TQA) processes and tools that are suitable for MT tasks. The following sections will be discussing the two main methods used for TQA: manual quality metrics—which will be discussed in Section 3.2.1.—and automatic quality metrics—discussed in Section 3.2.2..

3.2.1. Manual Quality Metrics

"Traditionally translation quality is evaluated by bilingual reviewers, who examine source and target texts to determine whether the translation meets requirements." (Lommel et al., 2014), however, this approach was very subjective since no metrics or tools were used to determine the quality of the translation, and practice varied from one LSP (Language Service Provider) to another (Lommel, 2018). For this reason, LSPs tried to minimize the source of inconsistency by implementing translation score-cards (typically spreadsheets) where the number of errors were counted and used to generate percentages that represented the quality of the translation (0-100%) (Lommel, 2018). Nevertheless, this practice was still not standardized and varied between LSPs, only giving the illusion of objectivity (Lommel, 2018).

The 1990s saw two efforts at standardization that, while quite different, both consisted of lists of error types:

- The first, the Localization Industry Standards Association (LISA) QA Model was first released as a spreadsheet[9] of error types that were categorized as minor, major or critical (Castilho et al., 2018). Each segment of text that was translated received a score based on the error types it contained and after all the scores were calculated, it would lead to

[9] Later on, the LISA QA Model was also released as a stand-alone software.

an overall score of the translated text as a whole, which would determine if the translation passed or failed (Castilho et al., 2018). Although this model was developed specifically for the translation of software documentation and User Interfaces, it was seen as a one-size-fits-all model and implementers used it for other content types (Lommel, 2018).

- The second, SAE J2450, was a score-card metric developed specifically for assessing the quality of translations of automotive documentation and contained six error types and two severity levels (Lommel, 2018).

The lack of error standardization across these models made it difficult to compare translation quality. Moreover, the widespread adoption of MT brought new, flexible quality requirements for which these models were not fit. LISA started developing a plan for a new, flexible standardized model. However, in 2011 it was dissolved and hopes for a new model were scrapped (Lommel, 2018). After the closure of LISA, two groups began working on TQA models:

The first one, the Translation Automation User Society (TAUS), developed the Dynamic Quality Framework (DQF) model, which proposed resolving quality issues before the translation process began, instead of posteriorly (Castilho et al., 2018). To create their error typology, TAUS gathered information on what was relevant for LSPs needs, focusing on solving specific issues instead of trying to address every possible need (Lommel, 2018). The DQF typology consisted of six error types (each one having a number of subtypes): accuracy, linguistic, terminology, style, country standards and layout; four categories to mark issues that were not considered errors: query implementation, client edit, repeat and kudos[10]; and four severities (Lommel, 2018).

The second one, the QTLaunchPad project, led by the German Research Centre for Artificial Intelligence (DFKI) created the Multidimensional Quality Metrics (MQM), a model that was developed by continuing the work of LISA and consisted of an "extensive translation error typology for use in detailed analysis of human and machine translation" (Lommel, 2018). MQM consists of a set of totally adaptable error types, that allows each user to create an MQM-based error typology that best fits their needs (Lommel, 2018). MQM is a highly hierarchical model that extends up to four layers of error types, each one becoming more specific with every layer. MQM has over 100 error types in total, with eight primary branches at the top of the

[10] For a detailed explanation of each error type and category, see Lommel (2018).

hierarchy: accuracy, design, fluency, internationalization, locale conventions, style, terminology, and verity; each of these branches out into several daughter error types, which are divided into granddaughter error types[11] (Lommel, 2018). Additionally, MQM has four severity levels: Null, Minor, Major and Critical. The Null severity is used to mark issues that are not considered errors and, as such, does not have a value when calculating the MQM score (Lommel, 2018). Minor errors are worth one point, whereas major errors are worth five points, and critical errors are worth ten points. These points are deducted from a total score of 100, yielding a MQM score calculated using the formula below:

=100-SUM((1*MINOR)+(5*MAJOR)+(10*CRITICAL))/#Words*100

Each user can choose the weight of each severity, although it should be noted that "doing so impedes comparison with scores generated using the default values" (Lommel, 2018).

Although TAUS DQF and MQM were being developed separately, in 2014 the two initiatives decided to create an integration of the two models. As a result, TAUS and DFKI submitted an application for the follow-up project QT21[12], which included a plan for the integration of the two (Lommel, 2018). To integrate both models, changes were made by both TAUS and DFKI, which resulted in the DQF-MQM integrated error typology, which, despite including fewer error types than the original MQM typology, continues to be customizable (Lommel, 2018).

3.2.2. Automatic Quality Metrics

Despite the fact that manual quality metrics might give a more complete picture of the quality of a translated text, the use of automatic quality metrics—either on their own or in conjunction with manual quality metrics—became popular, because manual quality metrics require more time and can be costly.

Bilingual Evaluation Understudy (BLEU), established by the IBM group in 2002, was the first automatic metric to be created (Hutchins, 2003). The developers of BLEU believe that "the closer a machine translation is to a professional human translation, the better it is" (Papineni et al., 2002), which is why this metric calculates the quality score of MT output by

[11] The full hierarchy and list of issues is available at http://www.qt21.eu/mqm-definition/issues-list.html
[12] For more information on the QT21 project, check: https://o.taus.net/qt21-project#dqf-qt21

cross-referencing it with previous translations produced by humans (Papineni et al., 2002). Thus, Papineni et al. (2002) claim that "our MT evaluation system requires two ingredients: 1. A numerical 'translation closeness' metric; 2. A corpus of good quality human reference translations". Although BLEU does show good correlation with human judgment, its judgment might still not be reliable due to the fact that, by judging MT output based on how closely it resembles the chosen reference translations, it may reject multiple valid translations because they do not resemble the reference translations (Castilho et al., 2018).

In 2005, Metric for Evaluation of Translation with Explicit Ordering (METEOR) is developed, a metric that resembles BLEU in the sense that it also cross-references the MT output with reference translations: "We describe METEOR, an automatic metric for machine translation evaluation that is based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translations" (Banerjee & Lavie, 2005). Despite their similarities, METEOR set out to overcome BLEU's fundamental flaws by permitting translations that contain synonyms, without restricting the MT output to reference translations. (Banerjee & Lavie, 2005).

In 2020, Unbabel presented "a neural framework for training multilingual machine translation evaluation models" (Rei et al., 2020) entitled Crosslingual Optimized Metric for Evaluation of Translation (COMET) to address the fact that "current metrics struggle to accurately correlate with human judgment at segment level and fail to adequately differentiate the highest performing MT systems" (Rei et al., 2020). As previously stated, NMT systems are the current state of the art, and due to their complexity, metrics such as BLEU and METEOR have proven insufficient to accurately determine the quality of MT output. Unlike previous metrics, COMET "exploits information from both the source input and a target-language reference translation" (Rei et al., 2020), instead of only using reference translations.

4. INTERNSHIP GOALS

As described in Section 2.1., Unbabel's annotation process relies on a community of annotators composed of professional linguists and translators. Annotations help identify quality problem areas around which Unbabel devises mitigating strategies to improve the quality of the translations it provides.

However, it is important to note that, while being the best tool to describe complex linguistic and translation phenomena, annotations are made by humans, and humans make mistakes. When Unbabel introduced a new, simplified error typology in March 2022, annotators had to learn a new set of errors, and go through an adjustment period. Since the introduction of this new typology, quality feedback has been sent to annotators approximately eighty times, regarding problems related to:

- Error Mislabeling – error types are incorrectly annotated as another error type
- Overannotation – words or phrases that should not be annotated are
- Severity – wrong severity is attributed to an error
- Underannotation – words or phrases that should be annotated are not
- Spanning – when selecting an error, the annotator selects a larger part of the text (that is not a part of the error)
- Interface – technical issues related to the annotation tool[13]

The following graph (Figure 7) shows which topics were mentioned the most when giving feedback to annotators, which allows us to see what annotators most have problems with.

[13] These issues are not relevant to the scope of the work developed during the internship so they will not be expanded upon.

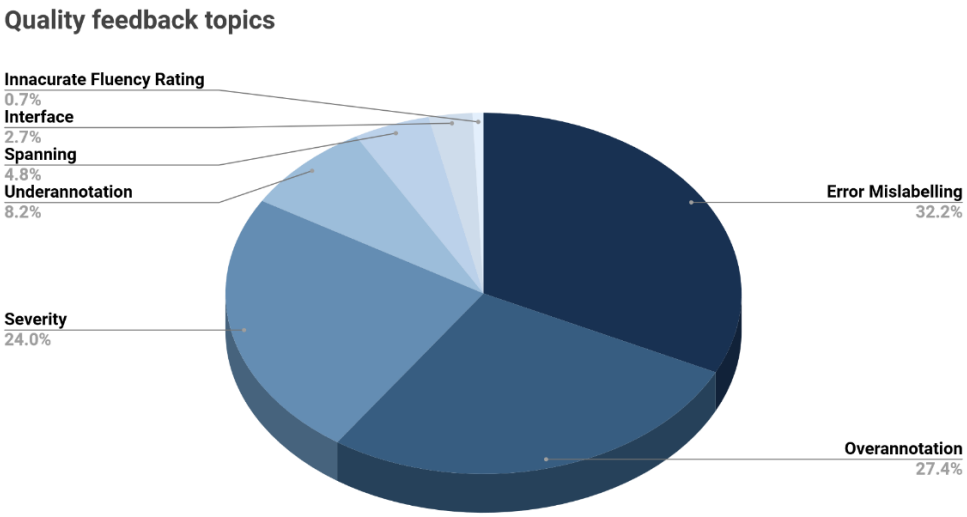


Figure 7 - Quality feedback topics

In light of these findings, we set out the research hypothesis that guides this work: by creating an annotation training, the reliability of annotations will improve. The objective was to provide annotators with the necessary knowledge and skills to reduce errors and enhance the quality of annotations.

These topics appear to cause annotators considerable uncertainty, which manifests itself in a number of incorrect annotations. This is because annotation is a complex task, and the annotation guidelines are extensive and information-dense. The following section describes the methodology followed to address this issue.

5. METHODOLOGY

A comprehensive methodology was developed with the objective of testing the hypothesis that training annotators will lead to improved annotations. First, Section 5.1. describes the process of creating an annotation training, with the aim of further determining which problems annotators had the most problems with and aid them in better understanding the annotation guidelines. Section 5.2. describes the process of creating an FAQ document to address annotators' most common doubts. Posteriorly, to assess whether the training had, in fact, been helpful, an observational study (Section 5.3.) was conducted to figure out if annotators made less annotation mistakes after taking the training.

5.1. Creating an Annotation Training

As discussed in the internship's goals (Section 4), the goal of this training was to help annotators better understand the errors they were making, while also letting them choose how much information they wanted to consume. To ensure a comfortable learning experience without overwhelming annotators with excessive information, the training begins by presenting a choice between a shorter or a longer version. This allows annotators to tailor the training to their preferences and learning style.

Throughout the training, annotators are also given the option to not re-read information they have already read, while still reading an explanation of why they got the question wrong. For example, if they have answered incorrectly to a Mistranslation vs. Unnatural Flow question and they get another question wrong in this category, they have the choice not to read extensively about the differences between Mistranslation and Unnatural Flow again (although they are encouraged to do so).

The training consists mostly of multiple-choice questions (except for three questions in the longer versions that are *yes or no* questions). Each topic has two or more questions to give the annotators a chance to learn from their mistakes.

As previously stated, there are two versions of the training available: a shorter version that can be completed in approximately 15 minutes and includes 10 questions, and a longer version that takes about 25 minutes to complete and comprises 17 questions. We suspected that the main issue annotators were having was with how extensive the guidelines were, therefore,

we thought it would be best to provide them with the alternative of something more easily digestible.

The shorter version of the training focuses on the most common error types, whereas the longer version takes a more structured approach, focusing on each parent category and also touching on the different severity levels.

Since the main goal of the short annotation training was to give annotators a quick way to understand difficult error types, it was not possible to cover Unbabel's error typology as a whole. Instead, the training focused on specific error types that proved to be more problematic, in some cases presented together with the errors for which they were more easily confused, to highlight when one or the other should be applied. In both versions these error types are covered:

- Mistranslation vs. Unnatural Flow
- Wrong Named Entity vs. Mistranslation, Untranslated and Locale Convention Issues
- Punctuation vs. Addition, Omission
- Punctuation vs. Locale Conventions
- Whitespace vs. Punctuation
- Term Not Applied vs. Wrong Term
- All of Locale Convention Issues
- Culture-Specific Reference vs. Wrong Language Variety
- Source Issue (Custom)

While the short version of the training only focuses on the above, in the longer version there are also questions about when to use other categories (Markup Tag, Do Not Translate, Lacks Creativity) and the four severities are also covered.

The training was sent to annotators who were found to be misusing the various error types and attributing the wrong severities.

Of the 16 annotators that were sent the training, only 10 completed it. The language pairs these annotators worked with are:

- en-pt-br (English to Brazilian Portuguese)
- en-pt (English to European Portuguese)
- en-de (English to German)
- en-nl (English to Dutch)

- en-cs (English to Czech)
- en-ja (English to Japanese)
- en-zh-CN (English to Simplified Chinese)

5.1.1. Annotator Feedback

At the end of both versions, there is a question with the intent of gathering their input on the training they took.

The qualitative feedback regarding the training content was overwhelmingly positive (independently of the version taken), with one annotator saying that the amount of content covered in the training was “just right” and another annotator claiming that they “would use this course as a constant aid while working on annotations. It is very helpful”.

However, there were some annotators that let us know how the training could have been better: three asked for “A deeper look into unnatural flow vs. mistranslation errors” and another said they would have liked to have seen “More examples of the items one answered wrong”.

5.1.2. Data Gathering

A data-oriented investigation was conducted to get a full understanding of the problem. By analyzing various jobs done by annotators and relying on specialist internal knowledge it was possible to discern the error types that were causing the most confusion and led to the most errors were:

(right category vs. wrong category)

- Mistranslation vs. Unnatural Flow
- Mistranslation vs. Grammar
- Mistranslation vs. Do Not Translate
- Mistranslation vs. Wrong Term
- Addition vs. Do Not Translate
- Term Not Applied vs. Wrong Term
- Unnatural Flow vs. Lacks Creativity
- Wrong Named Entity vs. Untranslated
- Wrong Named Entity vs. Mistranslation

- Wrong Named Entity vs. Locale Convention issues
- Locale Convention issues (e.g. Currency Format) vs. Whitespace
- Locale Convention issues vs. Word Order
- Do Not Translate vs. Company Style
- Wrong Language Variety vs. Culture-specific reference
- Punctuation vs. Addition
- Punctuation vs. Omission
- Markup Tag vs. Addition
- Markup Tag vs. Omission

5.2. Creating an FAQ for annotators

After creating the Annotation Training (short and long version), a document titled "FAQs for annotators" was also created to answer questions related to the problems most annotators seemed to have.

The document consisted of questions that were raised by annotators, as well as additional questions that we formulated based on the issues we had identified. We tried, to the best of our ability, to put these questions and answers into the categories that seem most relevant in relation to the Annotation Guidelines, and following the order set out in the typology. However, some questions belong in several categories and others in none. The category titles supplied in this document are meant to be seen more as a guide than a definitive categorization. Here is the structure of the document:

- Accuracy
- Linguistic Conventions
- Terminology
- Style
- Locale Conventions
- Design and Markup
- Custom (Source Issue)
- Other – this category is reserved for questions that do not fall under one of the parent categories from the company's error typology.

Since some of the questions can fall under more than one category, when this happens, at the end of the answer there is a small text saying 'This question and answer could equally relate to "[name of the category]"' that links the reader to the Annotation Guidelines.

Posteriorly, more questions and answers could, naturally, be added to this document.

5.3. Study Design (Observational study)

The study that was carried out during the internship was a Cross Sectional Observational Study[14]: "Observational studies are those where the researcher is documenting a naturally occurring relationship between the exposure and the outcome that he/she is studying. The researcher does not do any active intervention in any individual, and the exposure has already been decided naturally or by some other factor." (Ranganathan & Aggarwal, 2018).

The performances of particular annotators were monitored before and after taking the annotation training and the accuracy of their annotations was calculated by measuring what percentage of errors they annotated correctly (within each error category). The annotation assignments that were analyzed were chosen for reasons of convenience and no variables were manipulated in any way.

5.3.1. Participant selection

Ten annotators in seven LPs did the training. From those ten, due to internal knowledge of certain languages, only four language pairs—six annotators—were picked to be analyzed more carefully. Those six annotators are referred to as A1 to A6 throughout the study.

Annotator	Language Pair	Time working with Unbabel
A1	EN-PT-BR	> 2 years
A2	EN-PT-BR	< 6 months (From Lingo24)
A3	EN-PT-PT	> 1 years
A4	EN-DE	< 6 months (From Lingo24)
A5	EN-NL	< 6 months (From Lingo24)
A6	EN-NL	> 2 years

[14] Observational studies are commonly used in the area of translation as a way to understand certain processes.

A1, A3 and A6 had been with Unbabel for an extended period before taking the training; the other annotators, who had been with Unbabel for less than 6 months, had previously worked with Lingo24—a company acquired by Unbabel.

5.3.2. Data collection

Using Looker, a business intelligence tool, jobs annotated before and after the annotators completed the training were compared to determine the training's potential effectiveness. The jobs analyzed were a convenience sample, taken from two specific time frames: one before the training and one after the training.

Before beginning to evaluate their annotations, a timeframe was chosen to determine which jobs to study, and the MQM scores for these annotations were determined using the formula shown in Section 3.2.1.:

LP	Number of words	Minor	Major	Critical	MQM
en-pt-br	18298	161	183	21	92.97
en-pt	5695	190	5	0	96.22
en-de	7449	26	56	8	94.81
en-nl	7610	76	20	18	95.32

These scores were calculated prior to the annotation analysis because, depending on the errors made by the annotators, these values could be inaccurate, demonstrating how unreliable annotations can be.

6. ANALYSIS AND RESULTS

In this chapter, there will be a detailed analysis of the results found when analyzing annotation assignments of the six annotators that were chosen to participate in the observational study (participant selection is explained in Section 5.3.1.). The results will be divided into language pairs: EN-PT-BR (Section 6.1.), EN-PT (Section 6.2), EN-DE (Section 6.3.) and EN-NL (Section 6.4.); and in Section 6.5., the learning gathered from the study will be explained.

6.1. EN-PT-BR

6.1.1. A1

A1 only answered one question incorrectly during the Annotation training. Question number 8 (in the long version of the training) asked whether the "Lacks Creativity tag should be used whenever there is a part of the text that is not linguistically appealing", to which the annotator answered yes. However, Lacks Creativity should only be used when there is an express requirement from the customer, otherwise, Unnatural Flow should be used to mark an error that does not change the meaning of the sentence, but it is too literal a translation.

As can be seen in the chart (Figure 8), the error types A1 had the most problems with were Mistranslation and Punctuation, both incurring in more misannotations after the training. Hence, it is safe to assume that the training did not help this annotator with these specific error types, as they presented more misannotations after completing it. One possible explanation is that Mistranslation applies in different scenarios and overlaps with a few categories, so the examples presented may be only partially representative of the category as a whole. As for Punctuation, the training only presented it in relation to Locale Conventions, when there could be other causes for errors.

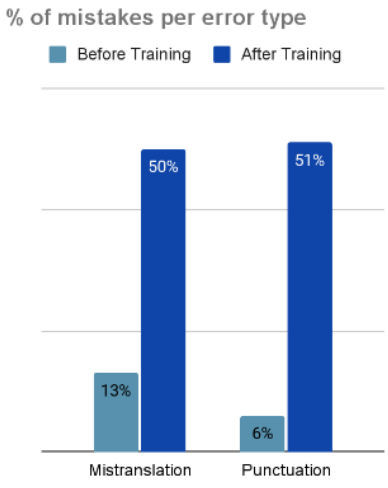


Figure 8 - % of mistakes made by A1

This annotator also proved to have a problem with using severities with no common thread, even attributing different severities to similar errors. An example of this can be found in Figures 9 and 10, where the annotator marked the first occurrence of the error as Major

(which is why it's highlighted in orange) and the second occurrence as Minor (which is highlighted in yellow). There is, however, no reason why the same error would be marked with two different severities. Still and all, the annotator seems to stop commiting severity mistakes after taking the training.

Figure 9 - Example of a currency format error (1)

Figure 10 - Example of a currency format error (2)

6.1.2. A2

A2 answered two questions incorrectly in the annotation training: question 3b (both versions of the training), about Mistranslation vs. Unnatural Flow[15] and question 11b (long version), about severities, having answered that a Major error was Critical.

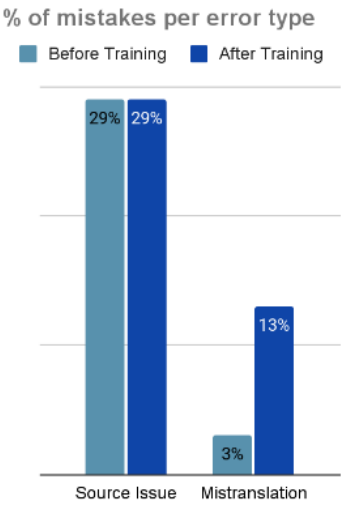


Figure 11 - % of mistakes made by A2

As can be seen in the chart (Figure 11), the error types A2 had the most problems with were Source Issue and Mistranslation. In terms of Source Issue, the annotator did not improve after taking the training and while the percentage of misannotations they produce using this error type is not that high, it shows that there are instances where the annotator uses Source Issue when they should not. It can be that they do not understand it fully. When it comes to the number of Mistranslation misannotations, these increased after the training and

[15] In the example shown in this question, "audience" was translated as "*audiência*", which is too literal a translation since in Portuguese "*audiência*" is used more as "hearing" (As in, court hearing).
Since this is an actual error and not just a matter of style, the correct answer would be Mistranslation and the annotator answered Unnatural Flow.

although it may not be a significant number, after analyzing these mistakes it became clear that this annotator has some issues understanding when to use Mistranslation or Unnatural Flow.

This annotator also shows some confusion around Omission after the training[16], seeing as they used it incorrectly one every three times (≈ 29%), as if something was omitted from the text when nothing was missing.

Although there is no data of A2 using Do Not Translate after taking the training, they did seem to struggle with it before since they used it multiple times (example in Figure 12[17]) when there was not a requirement from the client for something not to be translated (which is a requirement for the usage of this error type).

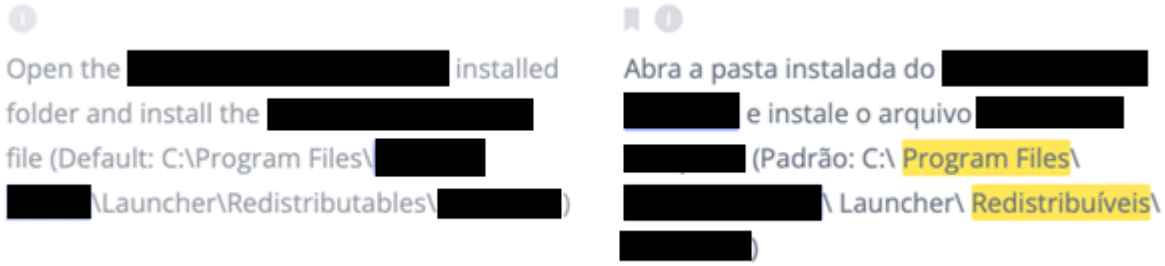


Figure 12 - Example of wrong annotation (not Do Not Translate)

Regarding severities, this annotator attributes the Critical severity a lot, even when it is not justifiable to do so. This is something that keeps occurring after taking the training.

6.2. EN-PT

6.2.1. A3

[16] There is no data of Omission being used before the training by this annotator.
[17] In this example, "*Redistribuíveis*" is marked as Do Not Translate and while this probably should not have been translated because it seems to be a computer path, Do Not Translate should only be used when there is a requirement from the client not to translate a specific term.

A3 only answered one question incorrectly during the Annotation training. Question 6b (both versions) asked whether a specific example[18] should be tagged as Wrong Term or Term Not Applied.

As can be seen in the chart (Figure 13), the error types A3 had the most problems with were Capitalization, Punctuation and Mistranslation. The annotator shows signs of confusion when it comes to Capitalization and Wrong Term since in instances like the one shown in Figure 14, they would tag the glossary term[19] as a Capitalization error since the word is capitalized. This remains a problem since after taking the training the annotator also uses Wrong Term incorrectly. A3's confusion with when to use Capitalization clearly remains after taking the training, which can also be said for Punctuation and Mistranslation.

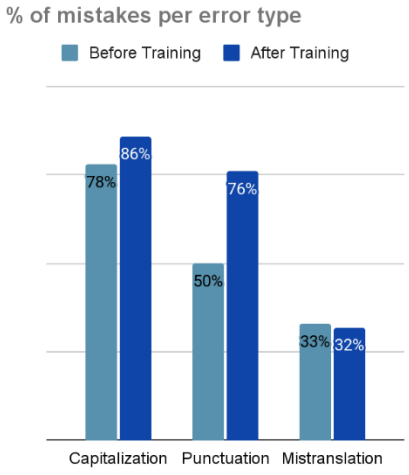


Figure 13 - % of mistakes made by A3

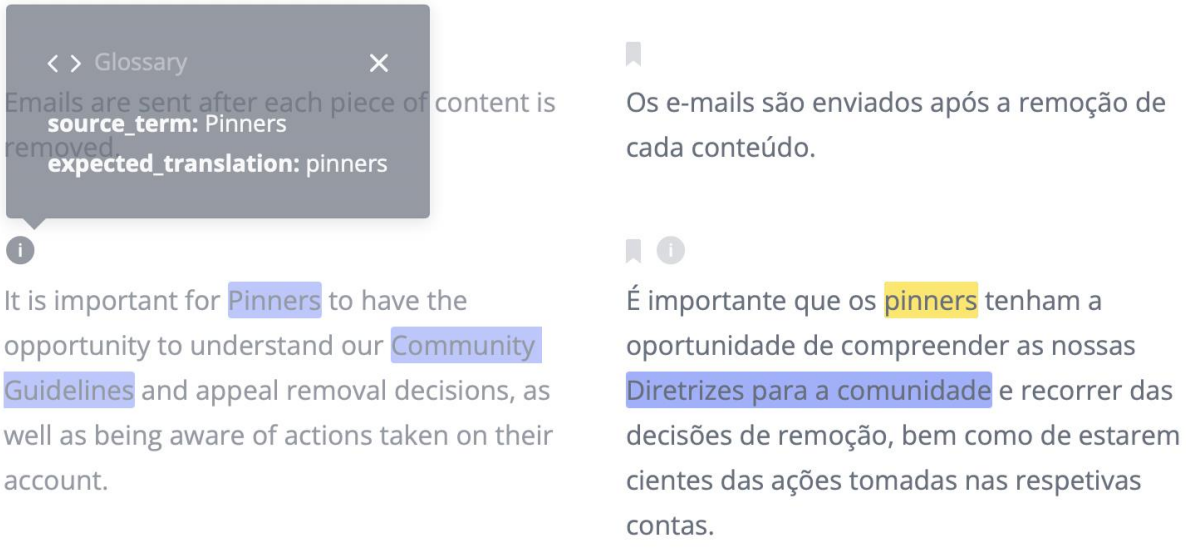


Figure 14 - Example of wrong annotation (not Capitalization)

There were also instances where the annotator used Source Issue incorrectly (≈ 20%) after taking the training and even though the number of mistakes made was not that high, it

[18] In this question we have a glossary term, meaning that if on the source text there is the word "profile", the expected translation for that term is "*Perfil*". The problem here is that in the term "*Perfil*" is written with a capital letter while in the sentence it should have been written in lowercase. This annotator answered that this was a Term Not Applied error, but the term is applied, it just does not fit the context, so the answer would be Wrong Term.

[19] Glossary term errors only fall under Term Not Applied or Wrong Term.

shows that there are cases where the annotator uses Source Issue when they should not, meaning that they do not completely understand when it should be used.

6.3. EN-DE

6.3.1. A4

A4 answered three questions incorrectly in the annotation training: question number 7 (long version), asking if the Do Not Translate tag[20] should be used whenever there is a part of the text that the annotator considers should not have been translated but was, question number 10 (long version), asking whether the Markup Tag should be used whenever there is a Markup[21] and question 11b (long version), about severities, having answered—like A2—that a Major error was Critical.

As can be seen in the chart (Figure 15), the error types A4 had the most problems with were Lacks Creativity and Wrong Named Entity. This annotator does not understand when to use Lacks Creativity at all and did not learn from the training since they keep using Lacks Creativity whenever they should use Unnatural Flow, seeing as Lacks Creativity should only be used when there is a specific client requirement. They also clearly have some confusion on what is considered a named entity, since Wrong Named Entity keeps being used whenever there is a markup in the text and a markup is not considered a named entity.

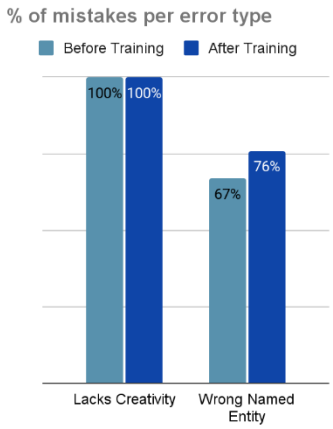


Figure 15 - % of mistakes made by A4

Before the training, the annotator classifies Whitespace errors as Spelling errors and attributes the Major severity (even if there is an error as small as a whitespace missing after a comma). Even though there is no data of the annotator using Spelling after taking the training—meaning that there is no way of knowing if they learned how to tag whitespace errors after taking the training—, A4 does keep attributing the Major severity to Whitespace errors.

[20] The Do Not Translate tag should only be used when something that the client specifically said should not be translated is, meaning that the answer to this question should be "no".
[21] The answer to this question is "No" because the Markup Tag should only be used when there is an error, if the Markup is well formed it should not be marked as an error.

6.4. EN-NL

6.4.1. A5

A5 only answered one question incorrectly during the Annotation training. This question (7a in short version; question 9a in long version) asked whether a specific example[22] should be tagged as Wrong Language Variety or Culture-specific Reference.

The case of this annotator is an interesting one, due to the errors they made. A5 strictly labels every error as Mistranslation, regardless of what type of error it is. To add to this, they attribute the Critical severity a lot, when it's not justifiable to do so. Even after taking the training and getting most questions right, it seems they either do not understand what they were meant to be doing or they just were not interested in learning.

For these reasons, A5 has been removed from the annotator pool.

6.4.2. A6

A6 answered three questions incorrectly in the annotation training: question number 7 (long version), about when Do Not Translate should be used—which we've covered previously in the A4 section—, question number 8 (long version), about when Lacks Creativity should be used—which we've covered previously in the A1 section—and question 11a, about severities, having answered that the severity that should only be reserved for Source Issue errors was the Major severity, when it is the Neutral severity.

As can be seen in the chart (Figure 16), the error type A6 had the most problems with was Untranslated, showing that they do not understand what a markup is, seeing as the instances when they incorrectly used the Untranslated tag was on markups—which should not be translated. Markups should only be tagged when they are incorrect (i.e. when the markup is not the same in the target text as it is in the source), using Markup Tag. The

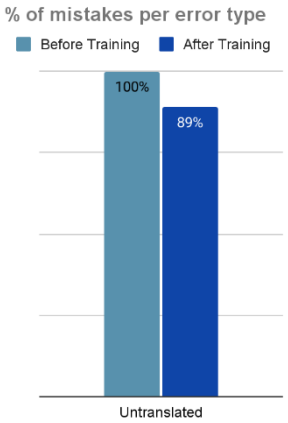


Figure 16 - % of mistakes made by A6

[22] Wrong Language Variety should be used when the language variety used is not the one requested, while Culture-specific Reference errors cover cases where the target text contains a culture-specific reference that's not appropriate or understandable to the intended target audience (such as jargon).
In this example, the translation is supposed to be in European Portuguese but the word "equipes" is Brazilian Portuguese, so the correct answer is Wrong Language Variety.

annotator keeps using Untranslated incorrectly after taking the training, meaning that they did not learn.

Although there is no data of the annotator using the Spelling tag after taking the training, they used it incorrectly several times before the training. Their misuse of the Spelling tag shows, once again, that the annotator does not understand what a markup is, since they use Spelling when the markup is incorrect—which, as we have seen, should be tagged as Markup Tag.

There were also no instances found of the annotator using the Source Issue tag after the training, however, before the training, they used it multiple times and always incorrectly. The Source Issue tag was used on its own, when it should always be used in combination with another error tag.

When it comes to severities, this annotator attributes the Critical severity a lot, even when it is not justifiable to do so. This is something that keeps occurring after taking the training.

6.5. Summary of results

We analyzed the annotations of six annotators in four language pairs. Except for one annotator, the percentage of misannotations due to misuse of the annotation category before the training was the same, or even higher, than after the training. When it came to severities, our results showed that even when annotators answered the questions correctly when taking the training, they still attributed higher severities than would be appropriate.

Despite the fact that only one (A5) of the six annotators whose annotations we examined completed the short version of the training, it appears that the same pattern was observed in both versions. The longer version of the training was expected to be more effective considering it was more structured and covered more topics; nevertheless, this assumption was proven incorrect.

Our hypothesis that by delivering training materials we could improve the reliability of annotations is then rejected. Most annotators continued to misannotate, even after taking the training and answering the questions correctly. In practice, the annotators had been annotating for some time and had some habits they didn't lose with the training. Some annotators, like A5, may simply be reluctant to learn. Nonetheless, this study did yield some important insights and,

consequently, ideas for future work that could effectively help annotators, which are discussed in Section 7.

7. CONCLUSIONS, LIMITATIONS AND FUTURE WORK

7.1. Conclusions

The goal of this report was to test the hypothesis that, by creating an annotation training, the reliability of annotations would improve. In light of this, an annotation training was developed and sent to annotators who were found to be missanotating on a regular basis. The plan was to analyze annotation assignments of annotators before and after taking the training. Even though we sent the annotation training to ten annotators working in seven languages, due to convenience reasons, we analyzed the annotation assignments of annotators who worked with language pairs for which there were internal language experts: EN-PT-BR, EN-PT, EN-DE and EN-NL. We also created an FAQ aimed at answering annotators' most pressing questions.

Unfortunately, after the analysis of the annotation assignments performed before and after taking the training, it became clear that neither a one-off annotation training or the FAQ were effective enough to improve the reliability of annotations made by the annotators that were present in the study. Among the potential reasons that could explain these results are:

a. The training was insufficient to encompass the complexity and/or nuances of the annotation task. For example, for clarity purposes, all sentences shown with errors in the training contained just one error, when in practice, real-life translations contain multiple, interacting errors.
b. The learning experience was disassociated from the actual tool annotators use to annotate.
c. The training was not adapted to individual styles, language phenomena, knowledge level or learning preferences. We learned from annotator feedback that some annotators wanted a more in-depth exploration of certain categories, or more examples. However, because the training was expected to be brief and not overwhelm the annotators with information, this proved to be a difficult task.
d. Lack of feedback loop. We didn't have the time to share detailed feedback with annotators to explain where they failed and why. This could have reinforced some knowledge, but due to time constraints associated with the duration of the internship, we'd need a follow up analysis that was out of scope of this internship.

Some additional improvements, more to do with user experience, that can contribute to have a more effective training are:

a. Incorporating the learning process into the annotation tool, to create a more integrated learning experience that could possibly result in a more effective learning experience.
b. In addition to educating annotators via trainings, making changes to the User Interface (UI) could contribute to lessening the misannotation problem. One possible issue is that when annotating, annotators do not consult the guidelines or, if they do, they find the error type name that most closely resembles the one they are looking for and do not bother reading others. Attempting to incorporate the guidelines into the UI could be beneficial (e.g. a pop-up that explains an error type when the annotator hovers or right-clicks on it [see Figures 17 and 18]). Making the information more accessible during the annotation process may encourage annotators to read the error type descriptions before selecting them. We also propose that, in some circumstances, when an annotator selects a sample of text, the annotation tool limits the error types the annotator could select (see Figures 19 and 20). For example, if an annotator selects a comma, the error type is likely Punctuation, thus, they would either only be allowed to select Punctuation or it would be highlighted. This would not work for all error types as some are more complex than others, but it would help reduce the number of annotation errors.

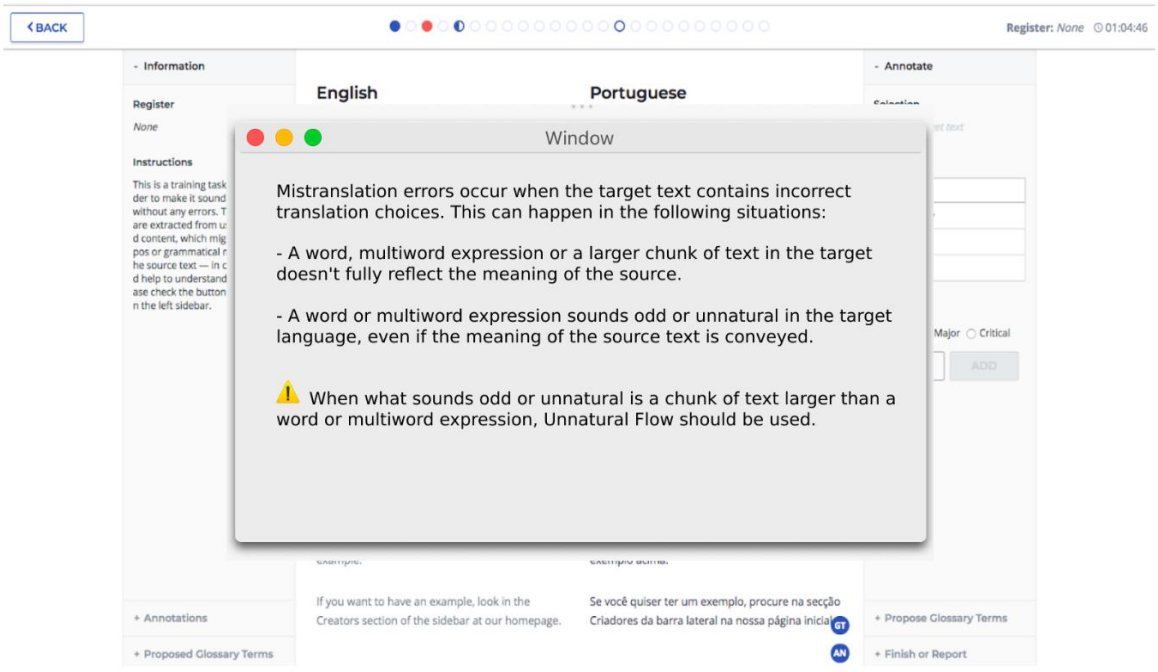


Figure 17 - Idea for Annotation Tool change (1)

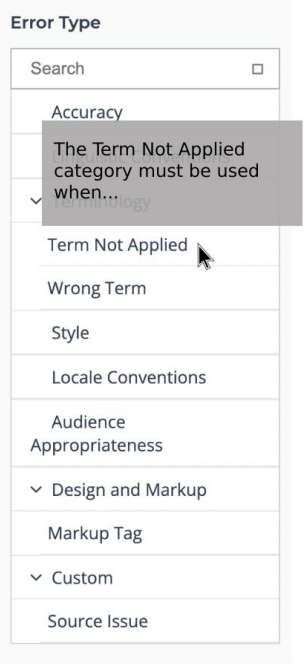


Figure 19 - Idea for Annotation Tool change (2)

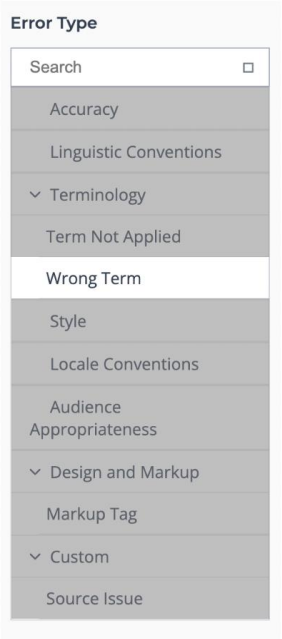


Figure 18 - Idea for Annotation Tool change (3)

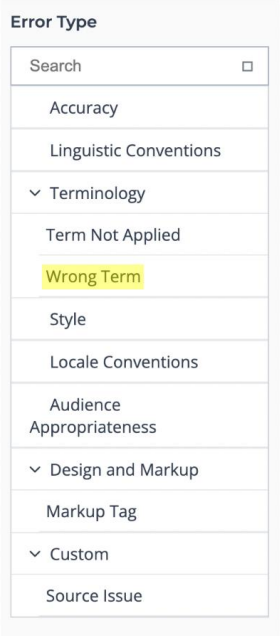


Figure 20 - Idea for Annotation Tool change (4)

7.2. Limitations

This was a one-off training initiative. A one-off training initiative may not be enough to achieve the learning goals set out at the beginning. Therefore, a more comprehensive and sustained approach may be needed. To ensure positive change from an annotator training initiative, it is essential to ensure that the training is:

- relevant, meaning that it needs to be adapted to the individual needs of annotators;
- delivered effectively, thus integrating it into the annotation tool would likely improve results, seeing as this way the learning would take place in the same environment where the annotators are going to work;
- followed up with reinforcement and support: a personalized approach/feedback-giving mechanism would likely help improve results.

In summary, annotation quality has to be approached from multiple angles, and ideally, has to be an effort sustained in time.

7.3. Future work

In this work, we've highlighted the challenges associated with training a freelance community of annotators via training materials. One future research avenue is exploring learning differences based on the amount of previous annotation experience. There is some anecdotal evidence that annotators who are new to the task may be more open to learn, since they have not had time yet to absorb potentially bad annotation habits. In fact, the longer training presented in this work is now used as an onboarding tool for new annotators at Unbabel.

Yet another avenue of research, would be letting annotators customize their learning experience by choosing which error categories or severities they want to reinforce, or the level of difficulty they want to tackle. We hope that future work will apply the learnings obtained in this study and explore various approaches to linguistic training in a freelance community.

8. BIBLIOGRAPHY

ALPAC, 1966. *Language and Computers in Translation and Linguistics*. https://nap.nationalacademies.org/resource/alpac_lm/ARC000005.pdf

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. https://aclanthology.org/W05-0909.pdf

Bar-Hillel, Y. (1960a). The Present Status of Automatic Translation of Languages. In *Advances in Computers* (Vol. 1, pp. 91–163). Elsevier. https://doi.org/10.1016/S0065-2458(08)60607-5

Castilho, S., et al. (2018). Approaches to Human and Machine Translation Quality Assessment. In *Translation Quality Assessment* (Vol. 1, pp. 9–38). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_2

Forcada, M. L. (2017). Making sense of neural machine translation. *Translation Spaces, 6(2)*, 291–305. https://doi.org/https://doi.org/10.1075/ts.6.2.06for

Gonçalves, M. (2021). *Analysis on the impact of the source text quality: Building a data driven typology*. https://repositorio.ul.pt/handle/10451/51178

Hutchins, J. (2003). *Machine Translation: A Concise History*. 1-21. https://xwiki.recursos.uoc.edu/wiki/mat00001ca/download/Research%20on%20Translation%20Technologies/Working%20with%20epub%20files%20using%20Python/WebHome/document3.pdf

Kenny, D. (2019). Machine Translation. In *The Routledge Handbook of Translation and Philosophy* (pp. 428–445). Routledge.

Koehn, P. (2020). *Neural Machine Translation* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108608480

Langlais, P., et al. (2005). *The Long-Term Forecast for Weather Bulletin Translation*. *19(1)*. 83–112.

Lommel, A., et al. (2014). Multidimensional quality metrics (mqm): a framework for declaring and describing translation quality metrics. *Revista Tradumàtica: Tecnologies de La Traducció*, (12), 455–463. https://doi.org/10.5565/rev/tradumatica.77

Lommel, A. (2018). Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In *Translation Quality Assessment* (Vol. 1, pp. 109–127). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_6

What is MQM?. MQM (Multidimensional Quality Metrics). (2022, August 29). https://themqm.org/

Papineni, K., et al. (2002). Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). In *BLEU: a Method for Automatic Evaluation of Machine Translation* (pp. 311–318). https://doi.org/10.3115/1073083.1073135

Paulo, M. (2022). *Analysis of context-aware Translation Memories: Part-of-Speech pattern distribution and gender neutral Translation Memories*. https://repositorio.ul.pt/handle/10451/56103

QT21 Project. TAUS. (n.d.). https://o.taus.net/qt21-project#dqf-qt21

Rei, R. (2020). Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). In *COMET: A Neural Framework for MT Evaluation* (pp. 2685–2702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Stein, D. (2018). Machine translation: Past, present and future. In *Language technologies for a multilingual Europe* (pp. 5–17). Language Science Press. https://doi.org/10.5281/zenodo.1291924

Silva, B. (2022). *Translation Error Annotation: Building an Annotation Module for East Asian Languages*. https://repositorio.ul.pt/handle/10451/56071

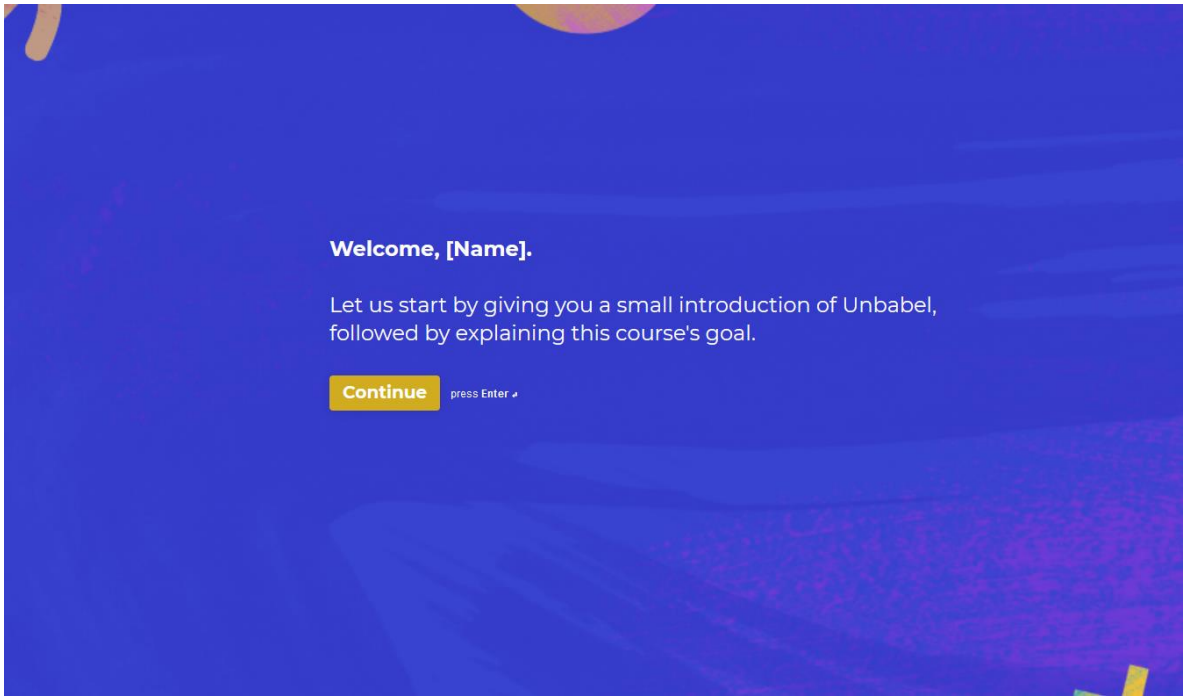
The MQM error typology. MQM (Multidimensional Quality Metrics). (2023, April 10). https://themqm.org/error-types-2/typology/

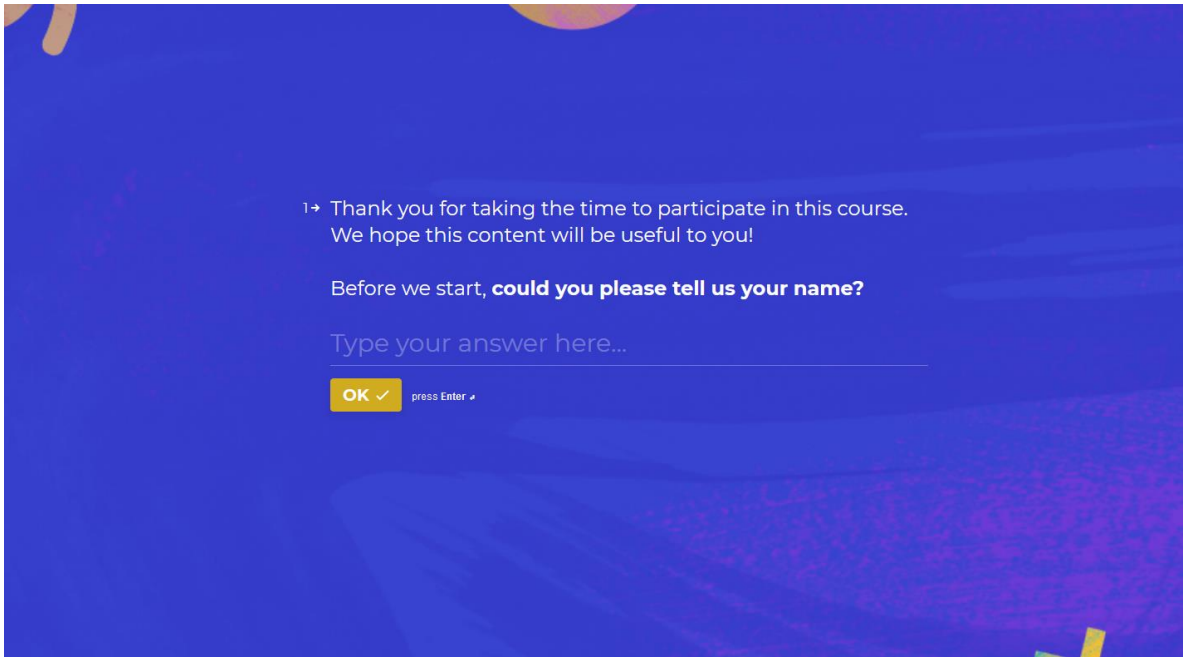
Unbabel launches Comet, blazing new trail for ultra-accurate machine translation. Unbabel. (2022, December 9). https://resources.unbabel.com/press-releases/unbabel-launches-comet-blazing-new-trail-for-ultra-accurate-machine-translation

9. LIST OF FIGURES

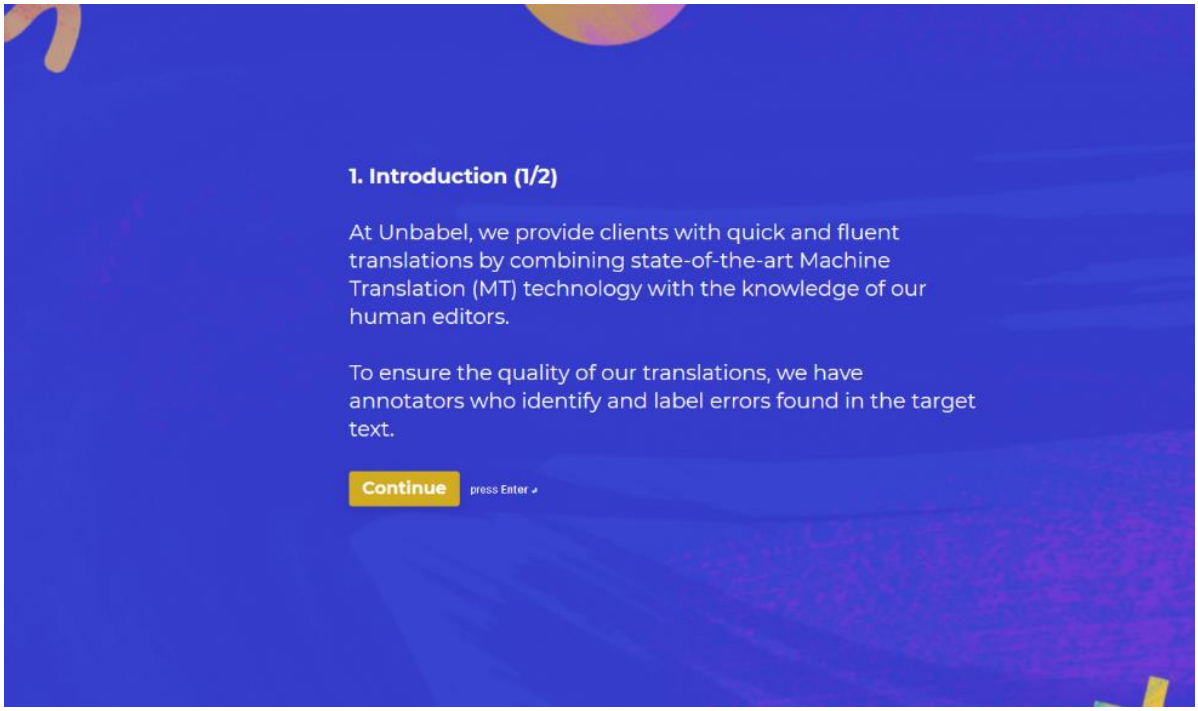
10. ANNEXES

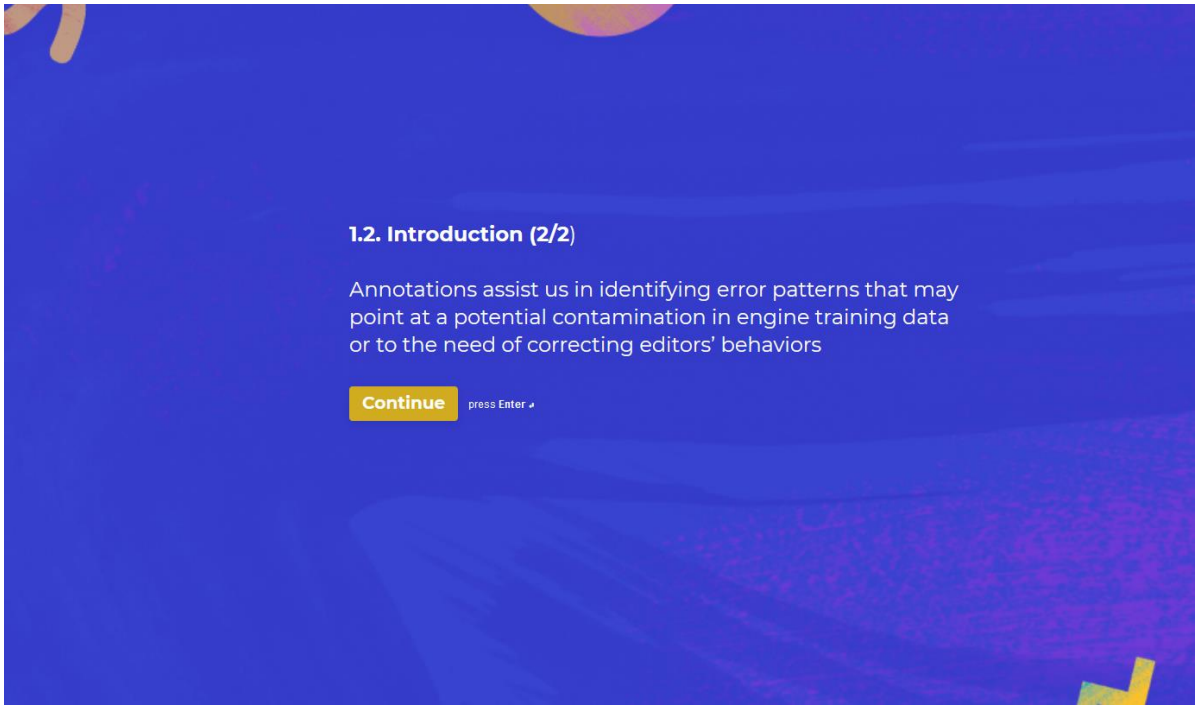
Annotation training (short version)[23]

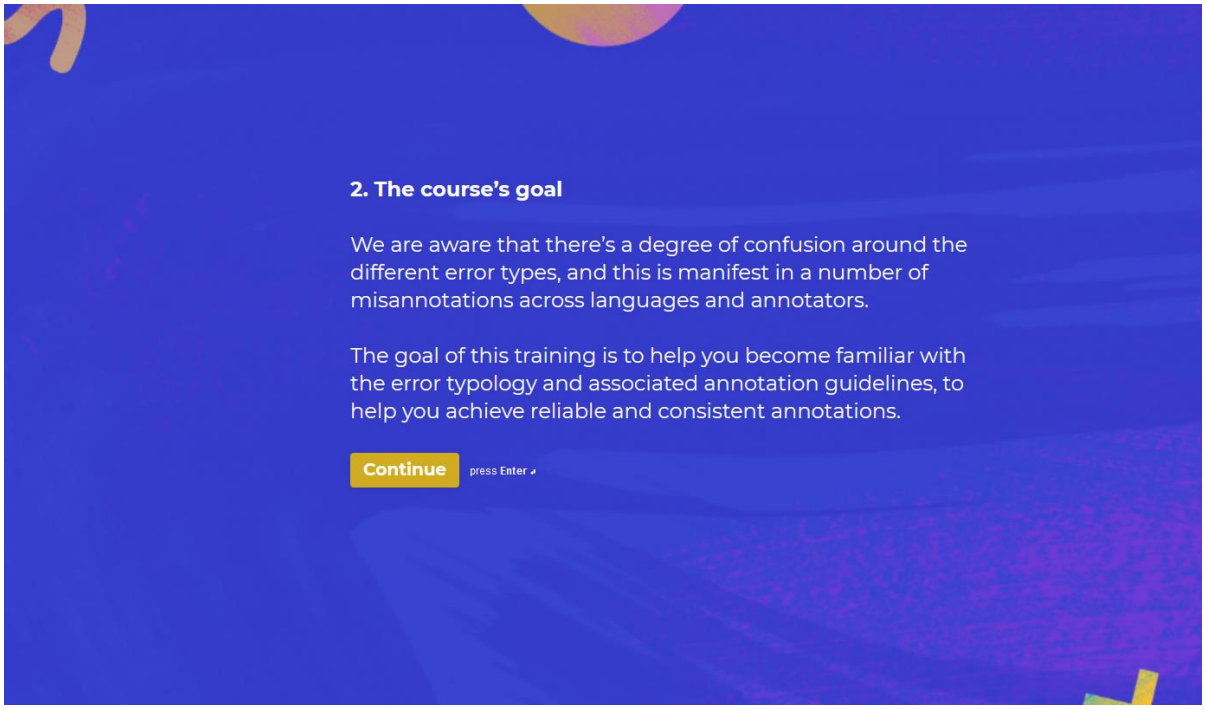


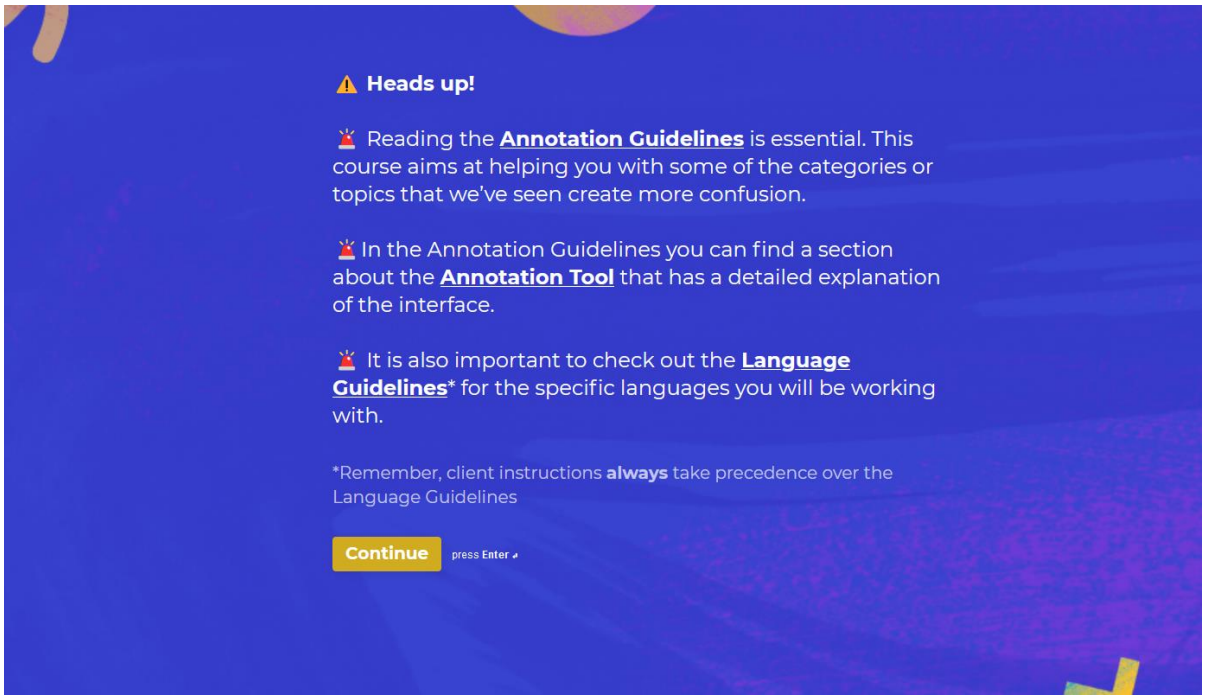


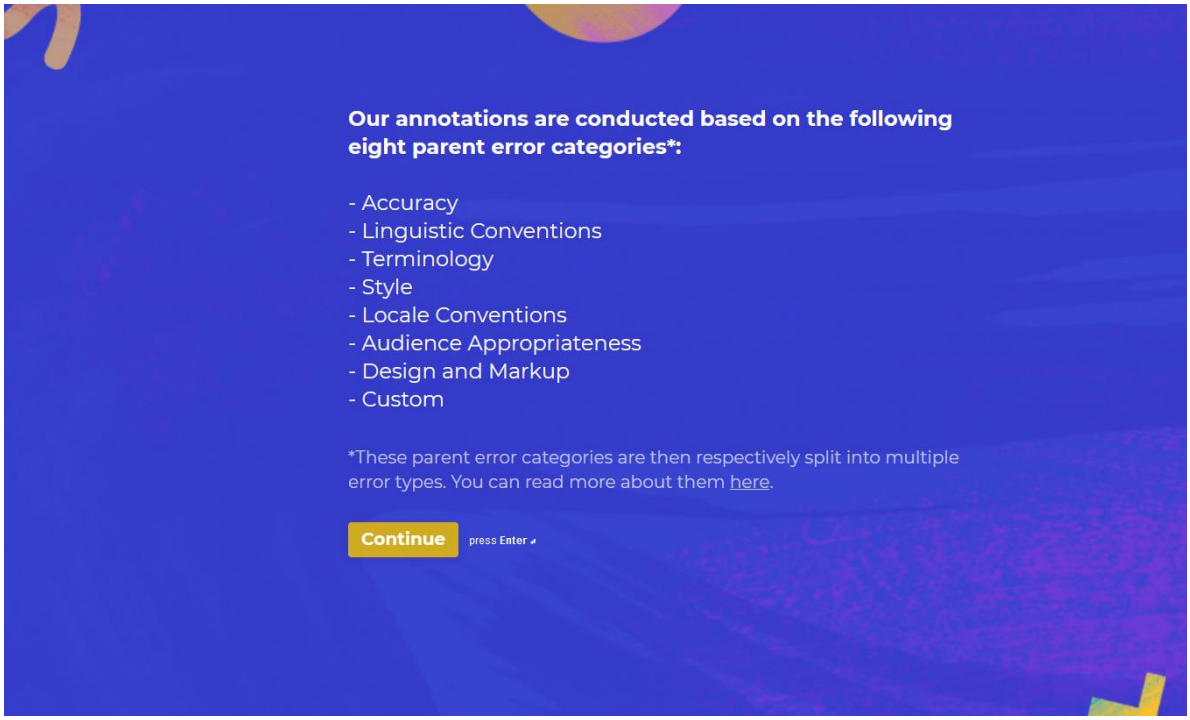
[23] The right answer for each question is selected. However, to demonstrate what annotators would see if they answered incorrectly, I answered every question incorrectly.

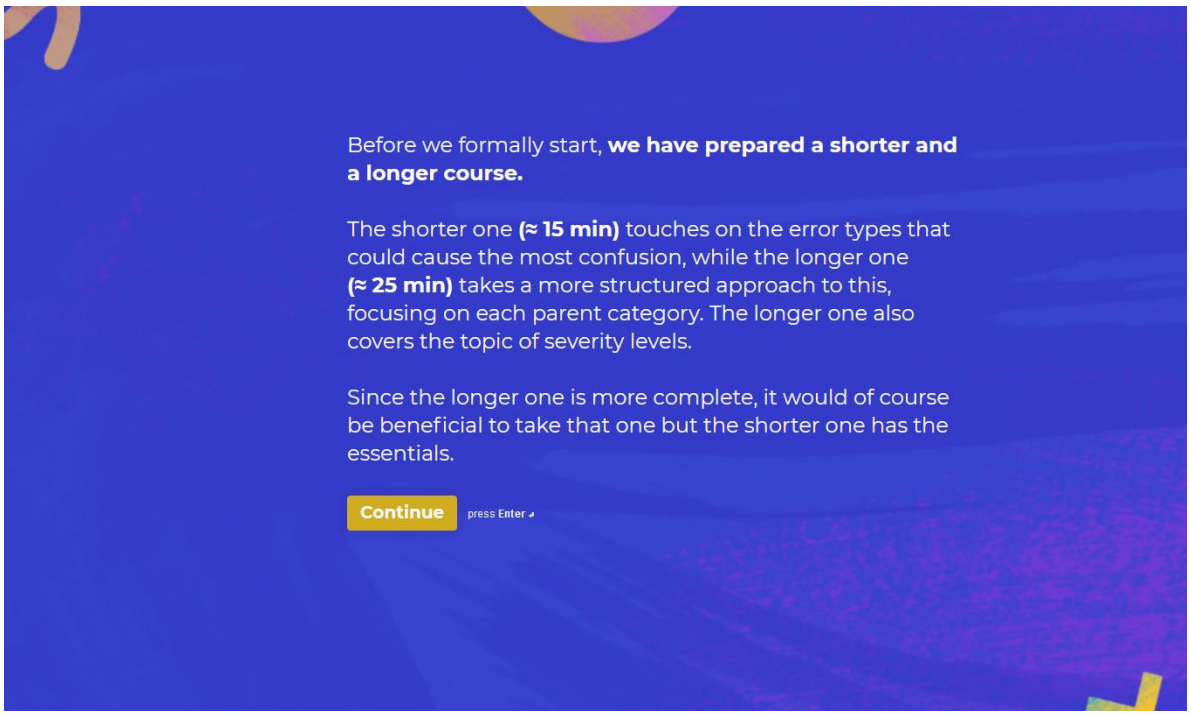

1. Introduction (1/2)
At Unbabel, we provide clients with quick and fluent translations by combining state-of-the-art Machine Translation (MT) technology with the knowledge of our human editors.
To ensure the quality of our translations, we have annotators who identify and label errors found in the target text.
Continue
press Enter

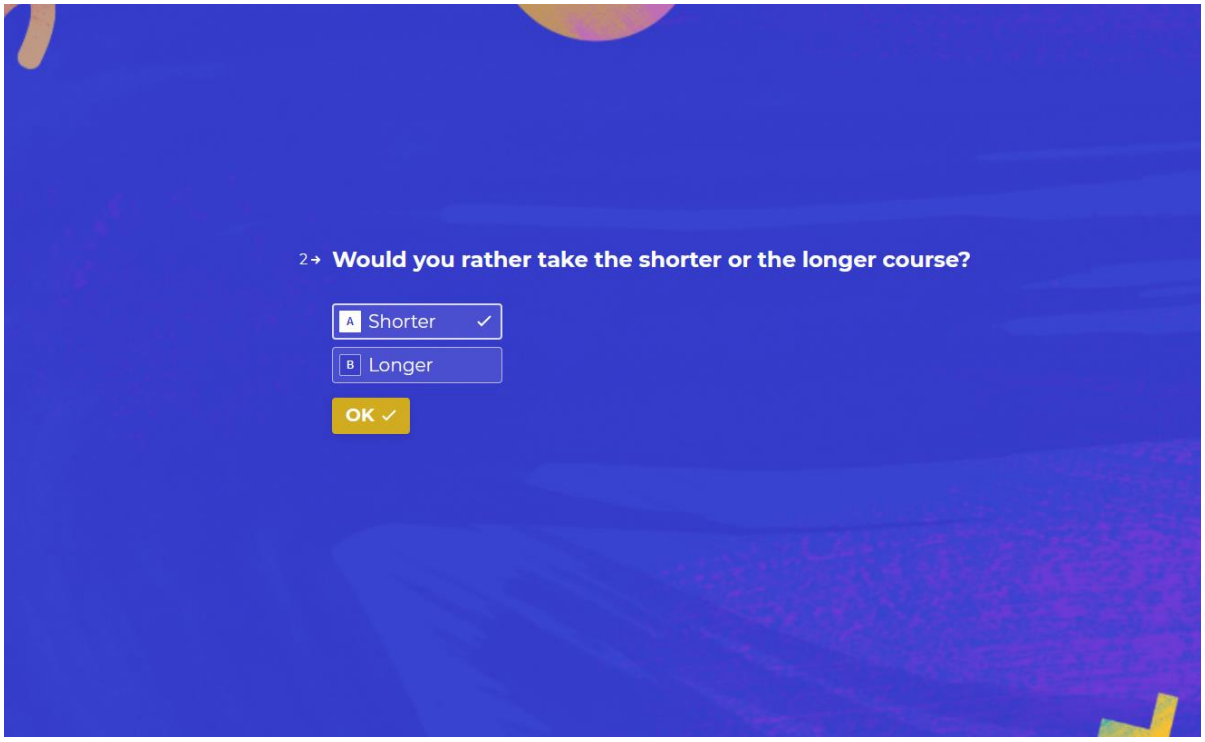

1.2. Introduction (2/2)
Annotations assist us in identifying error patterns that may point at a potential contamination in engine training data or to the need of correcting editors' behaviors
Continue
press Enter

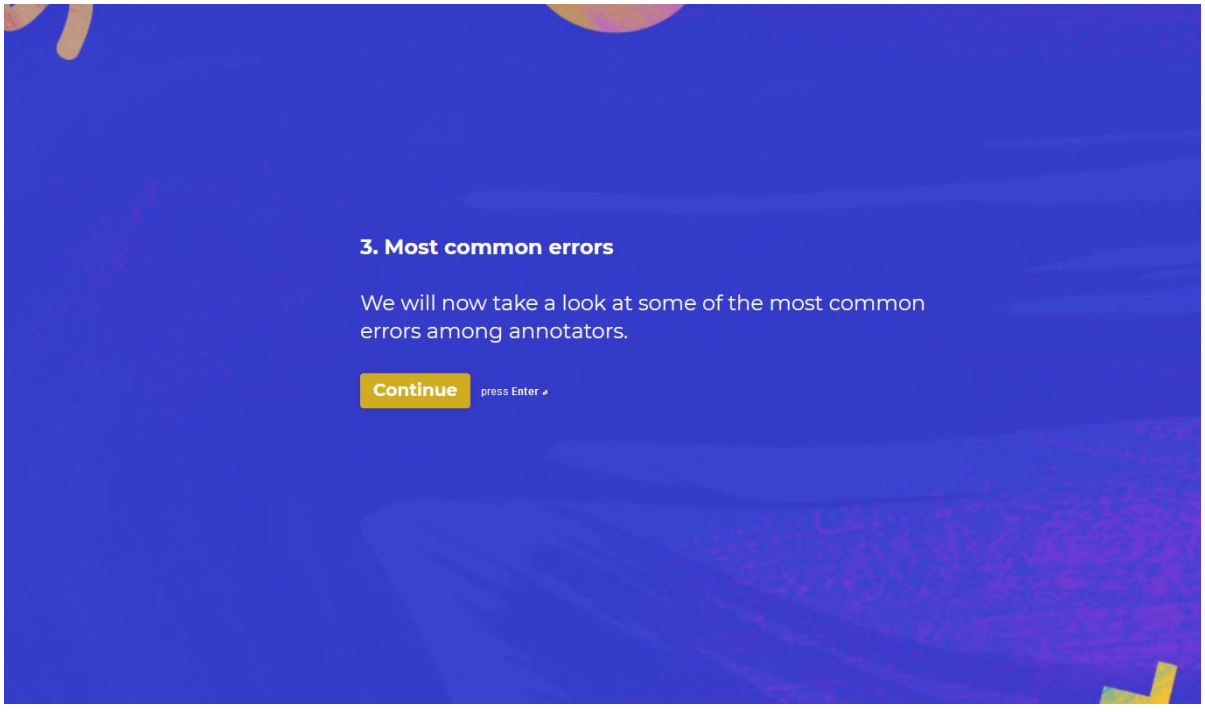

2. The course's goal
We are aware that there's a degree of confusion around the different error types, and this is manifest in a number of misannotations across languages and annotators.
The goal of this training is to help you become familiar with the error typology and associated annotation guidelines, to help you achieve reliable and consistent annotations.
Continue
press Enter

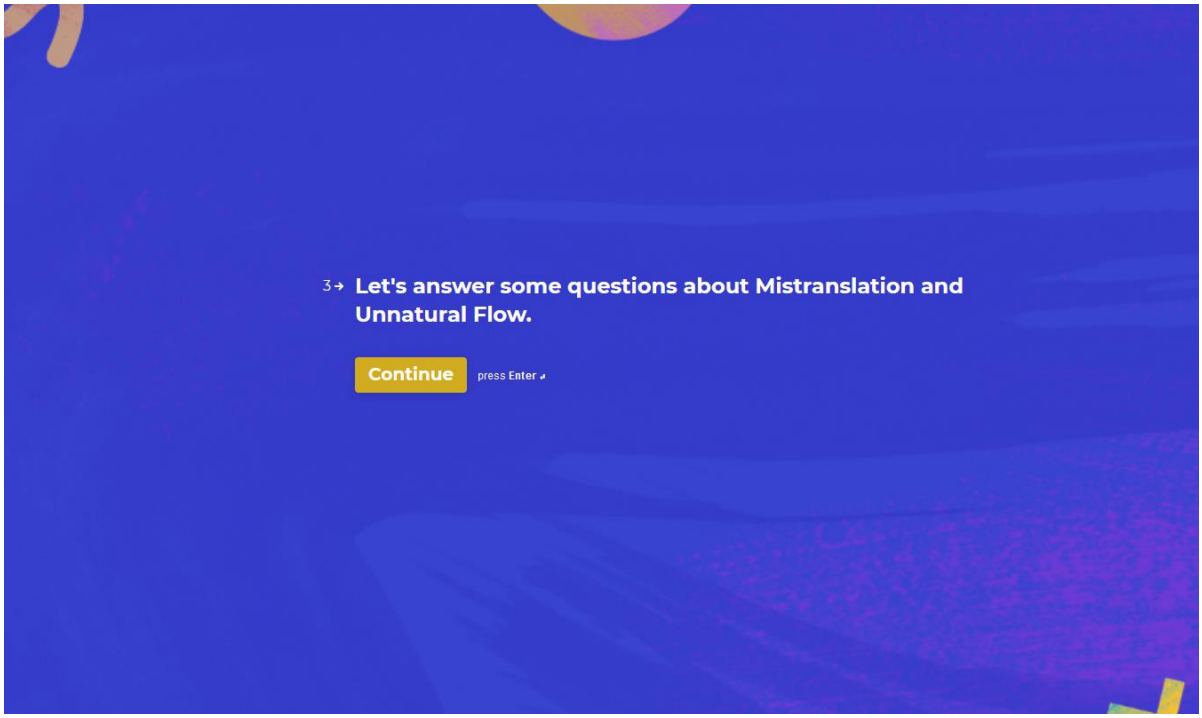

⚠ Heads up!
Reading the Annotation Guidelines is essential. This course aims at helping you with some of the categories or topics that we've seen create more confusion.
In the Annotation Guidelines you can find a section about the Annotation Tool that has a detailed explanation of the interface.
It is also important to check out the Language Guidelines* for the specific languages you will be working with.
*Remember, client instructions always take precedence over the Language Guidelines
Continue
press Enter

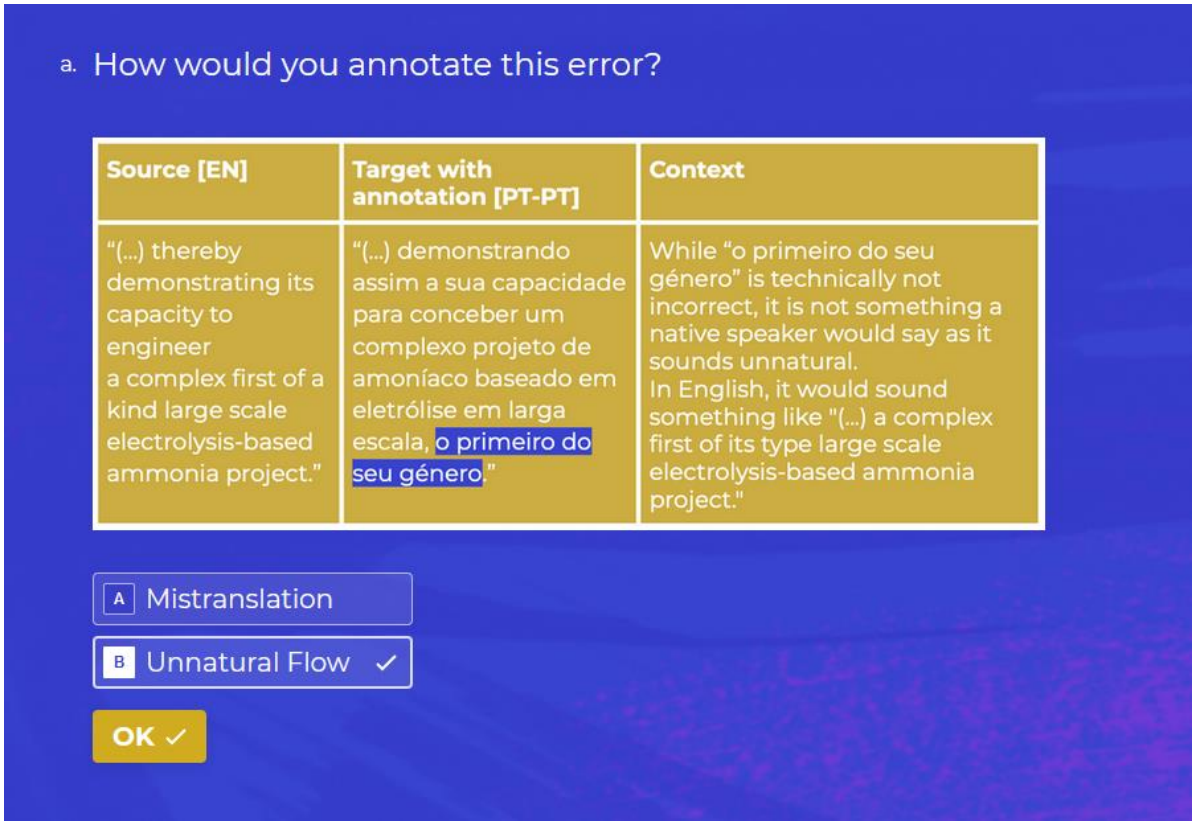

Our annotations are conducted based on the following eight parent error categories*:
- Accuracy
- Linguistic Conventions
- Terminology
- Style
- Locale Conventions
- Audience Appropriateness
- Design and Markup
- Custom
*These parent error categories are then respectively split into multiple error types. You can read more about them here.
Continue
press Enter

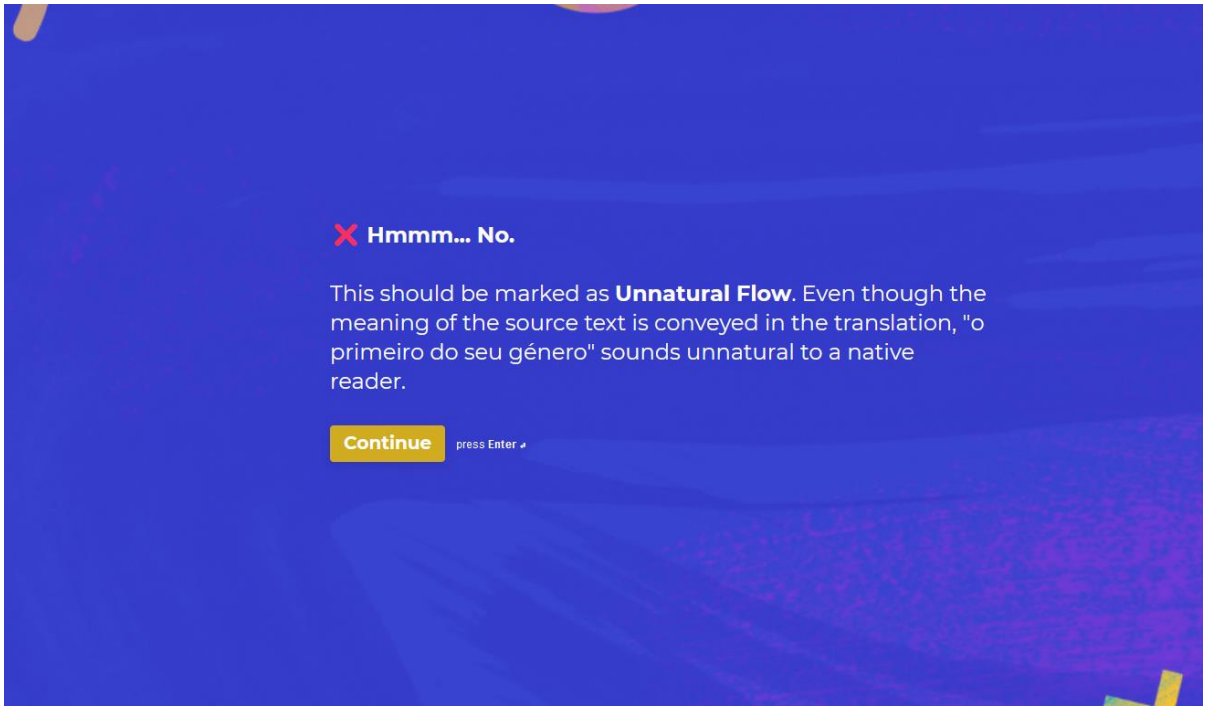

Before we formally start, we have prepared a shorter and a longer course.
The shorter one (≈ 15 min) touches on the error types that could cause the most confusion, while the longer one (≈ 25 min) takes a more structured approach to this, focusing on each parent category. The longer one also covers the topic of severity levels.
Since the longer one is more complete, it would of course be beneficial to take that one but the shorter one has the essentials.
Continue
press Enter

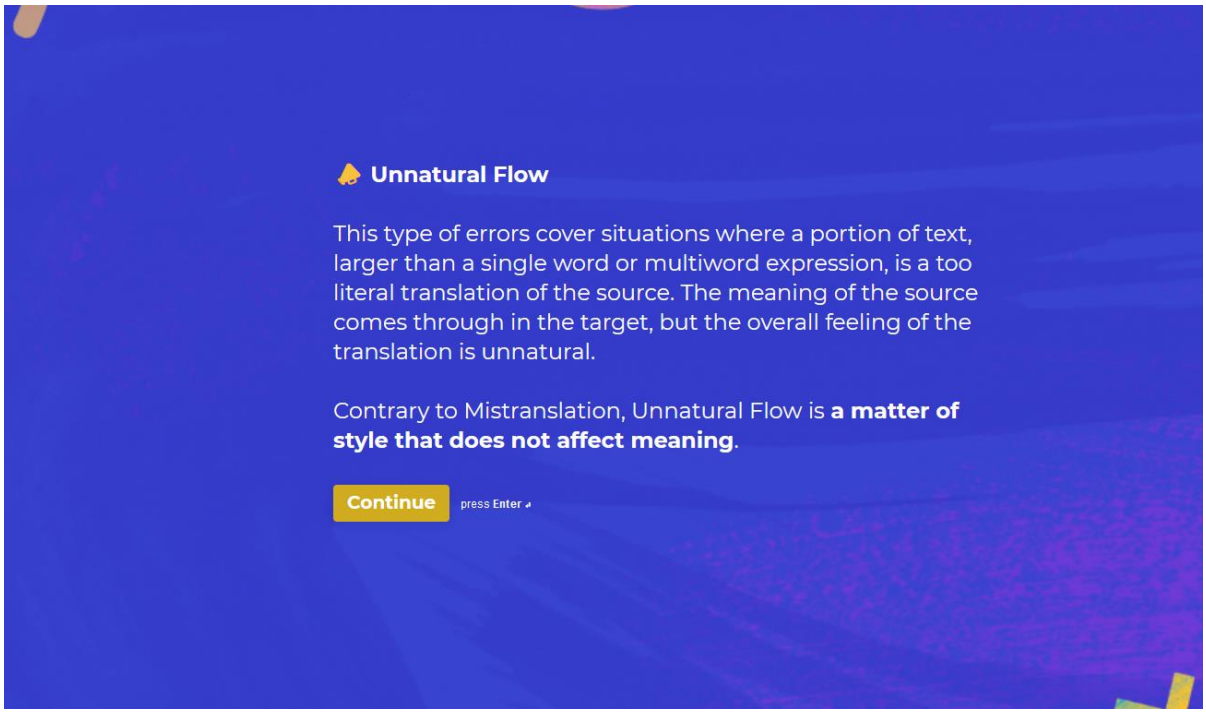

2→ Would you rather take the shorter or the longer course?
A Shorter ✓
B Longer
OK ✓

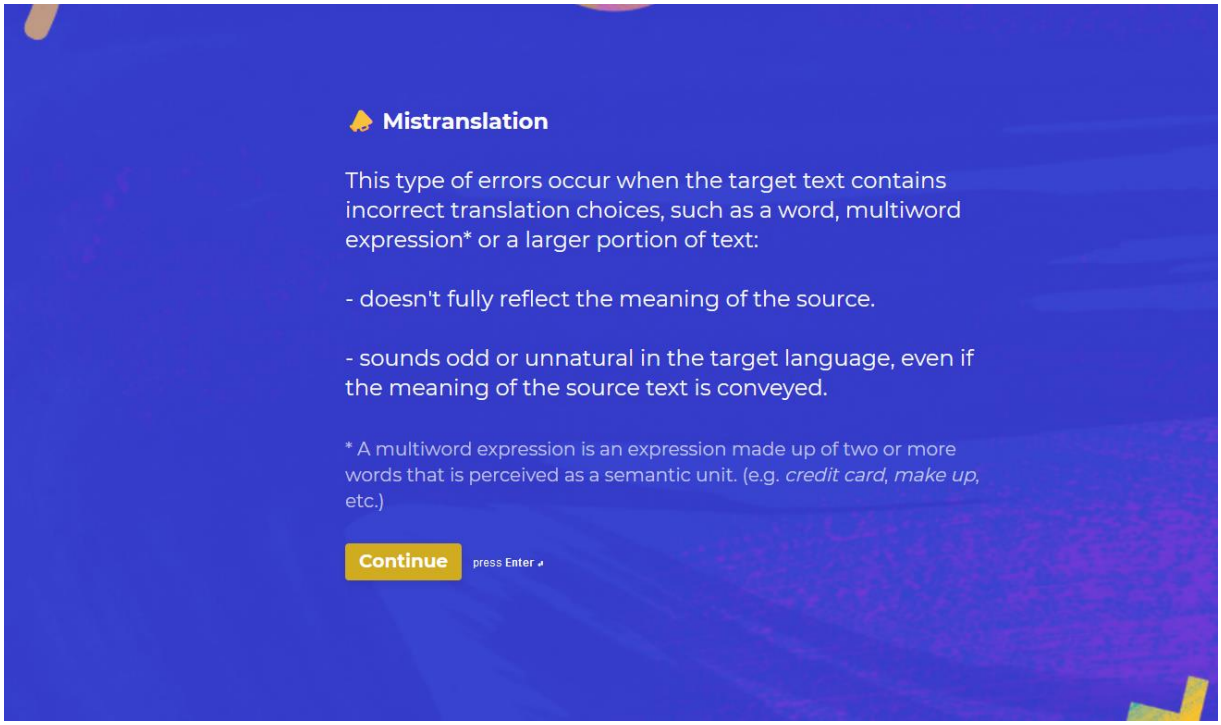

3. Most common errors
We will now take a look at some of the most common errors among annotators.
Continue
press Enter

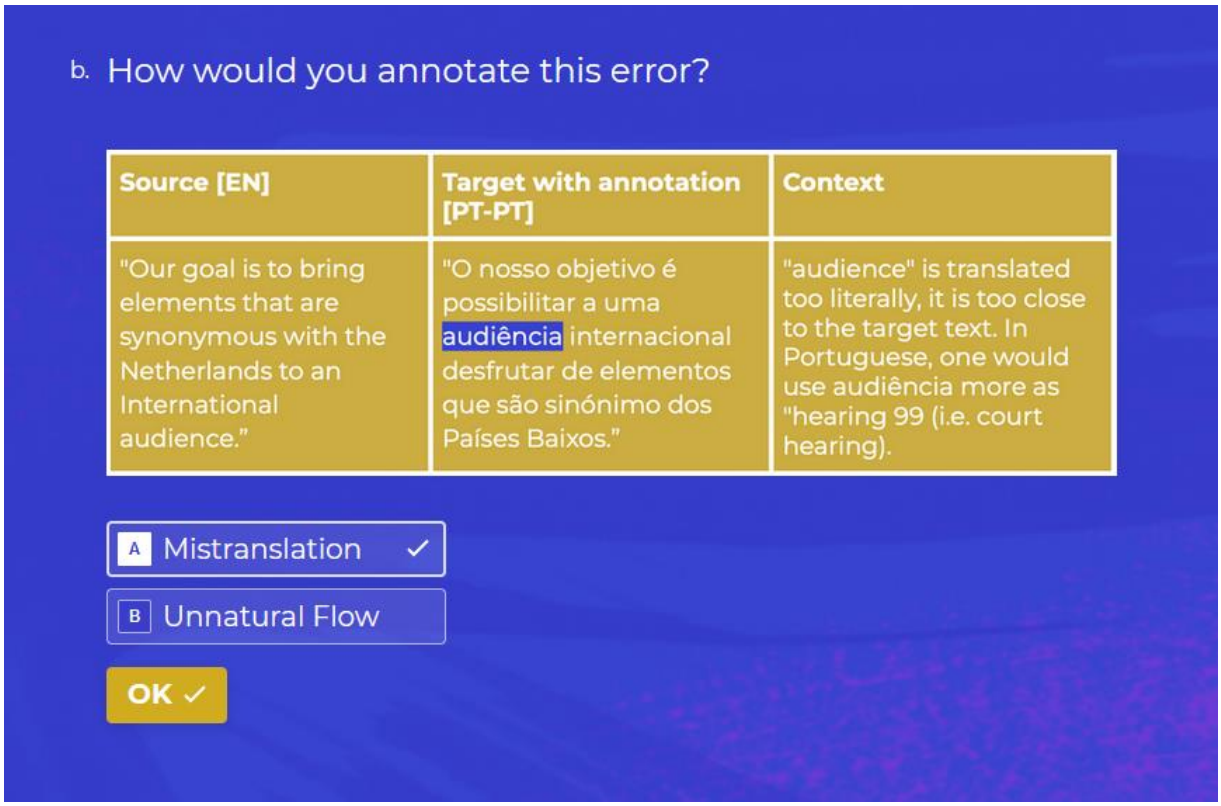

3→ Let's answer some questions about Mistranslation and Unnatural Flow.
Continue
press Enter

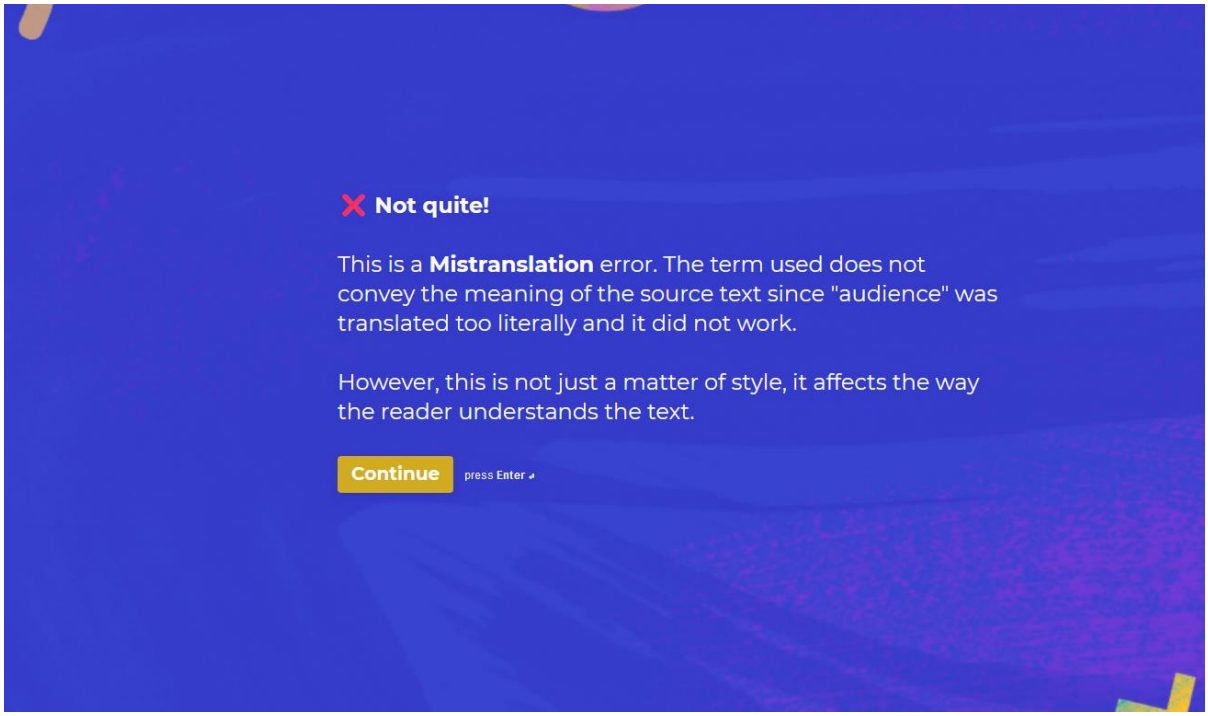

a. How would you annotate this error?
Source [EN]
Target with annotation [PT-PT]
Context
"(...) thereby demonstrating its capacity to engineer a complex first of a kind large scale electrolysis-based ammonia project."
"(...) demonstrando assim a sua capacidade para conceber um complexo projeto de amoníaco baseado em eletrólise em larga escala, o primeiro do seu género."
While "o primeiro do seu género" is technically not incorrect, it is not something a native speaker would say as it sounds unnatural.
In English, it would sound something like "(...) a complex first of its type large scale electrolysis-based ammonia project."
A Mistranslation
B Unnatural Flow ✓
OK ✓

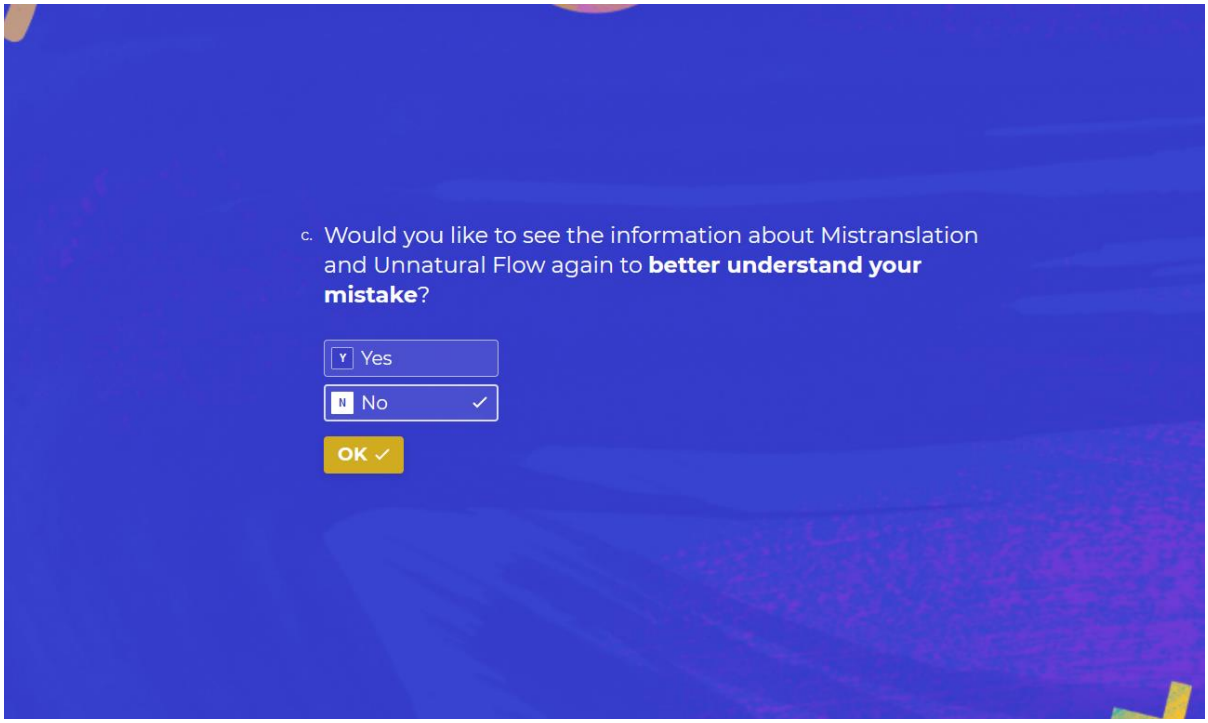

❌ Hmmm... No.
This should be marked as **Unnatural Flow**. Even though the meaning of the source text is conveyed in the translation, "o primeiro do seu género" sounds unnatural to a native reader.
Continue
press Enter

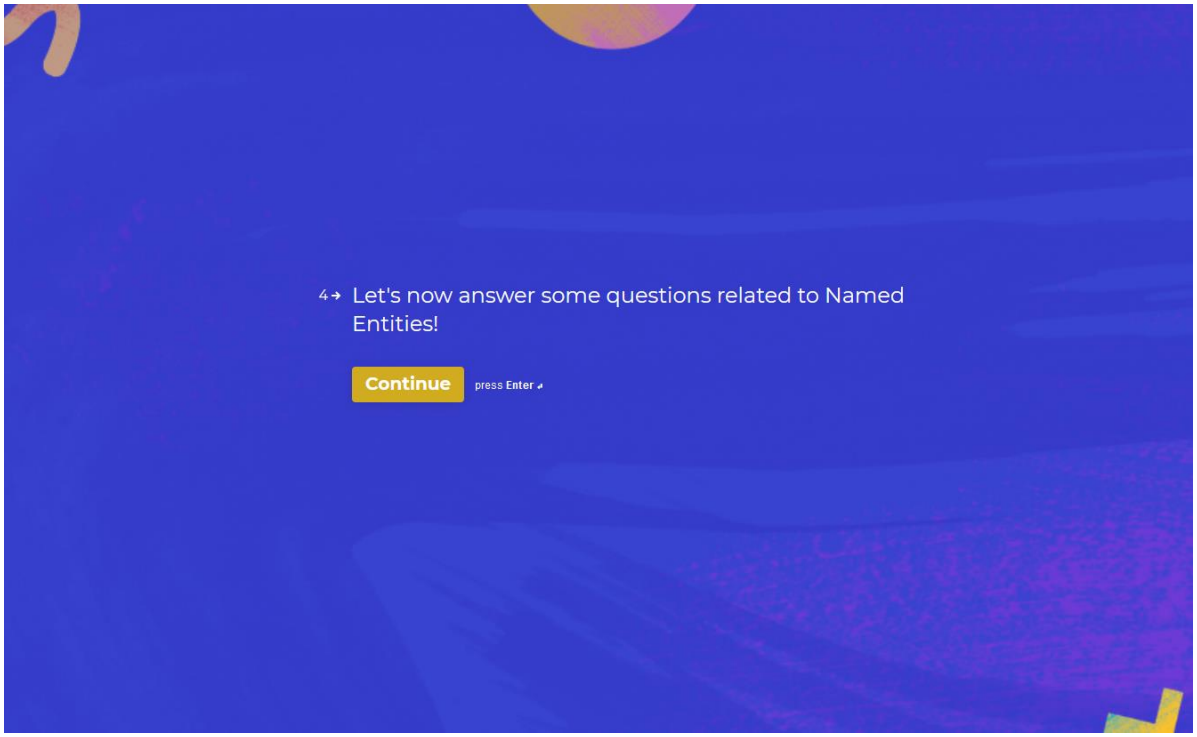

Unnatural Flow
This type of errors cover situations where a portion of text, larger than a single word or multiword expression, is a too literal translation of the source. The meaning of the source comes through in the target, but the overall feeling of the translation is unnatural.
Contrary to Mistranslation, Unnatural Flow is **a matter of style that does not affect meaning**.
Continue
press Enter

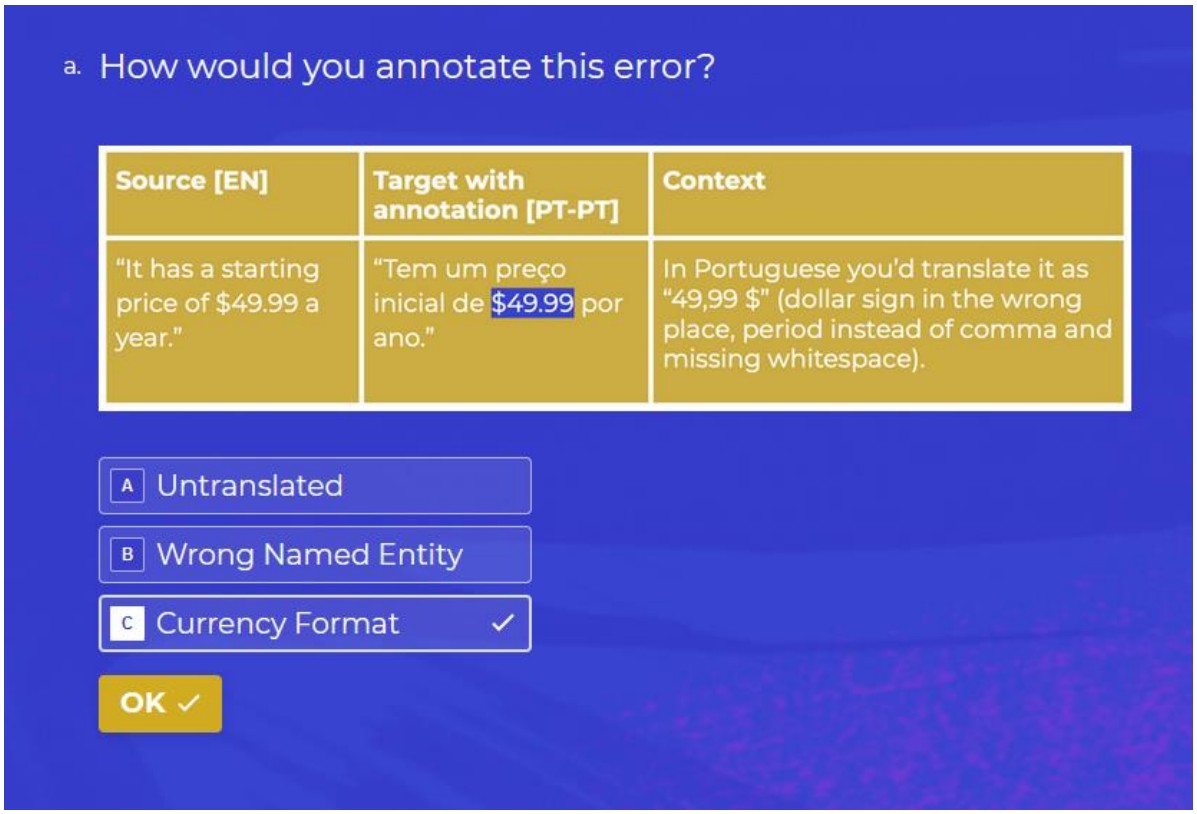

Mistranslation
This type of errors occur when the target text contains incorrect translation choices, such as a word, multiword expression* or a larger portion of text:
- doesn't fully reflect the meaning of the source.
- sounds odd or unnatural in the target language, even if the meaning of the source text is conveyed.
* A multiword expression is an expression made up of two or more words that is perceived as a semantic unit. (e.g. credit card, make up, etc.)
Continue
press Enter

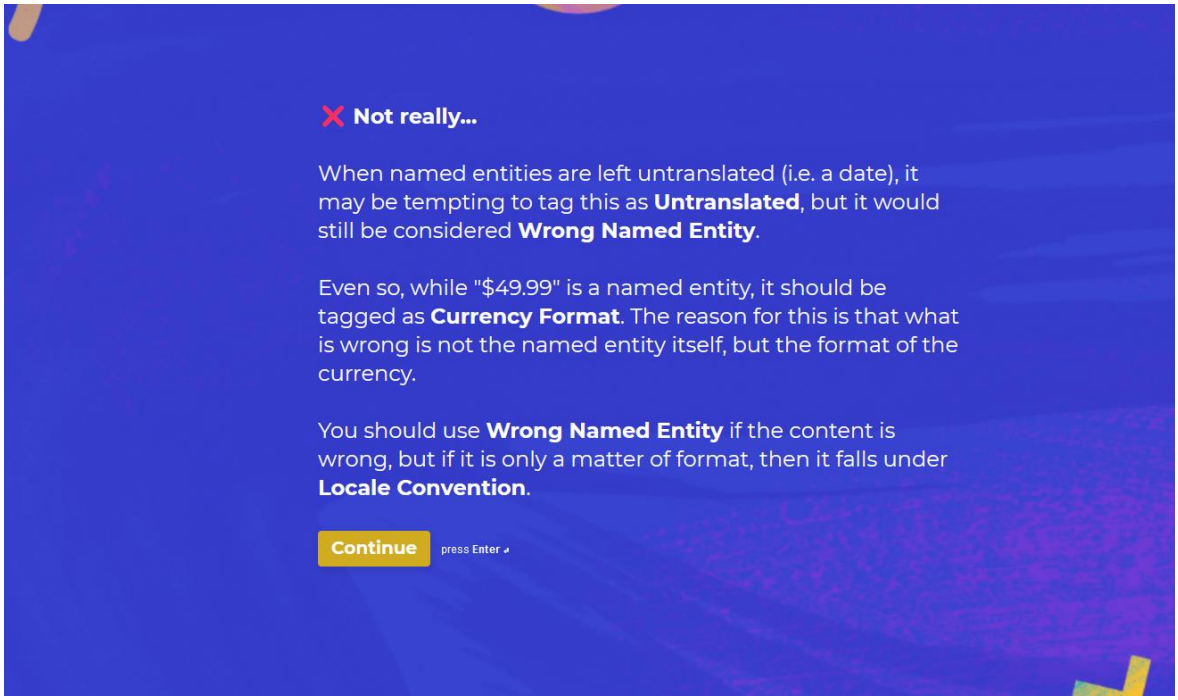

b. How would you annotate this error?
Source [EN]
Target with annotation [PT-PT]
Context
"Our goal is to bring elements that are synonymous with the Netherlands to an International audience."
"O nosso objetivo é possibilitar a uma audiência internacional desfrutar de elementos que são sinónimo dos Países Baixos."
"audience" is translated too literally, it is too close to the target text. In Portuguese, one would use audiência more as "hearing 99 (i.e. court hearing).
A Mistranslation
B Unnatural Flow
OK

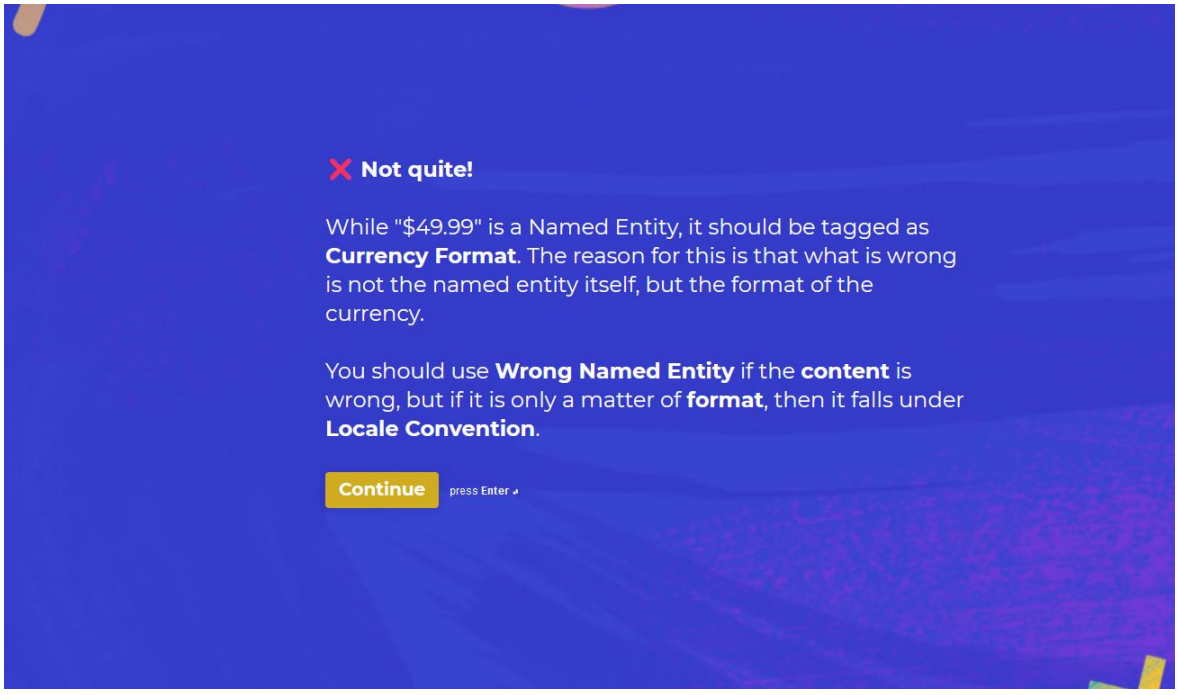

❌ Not quite!
This is a **Mistranslation** error. The term used does not convey the meaning of the source text since "audience" was translated too literally and it did not work.
However, this is not just a matter of style, it affects the way the reader understands the text.
Continue
press Enter

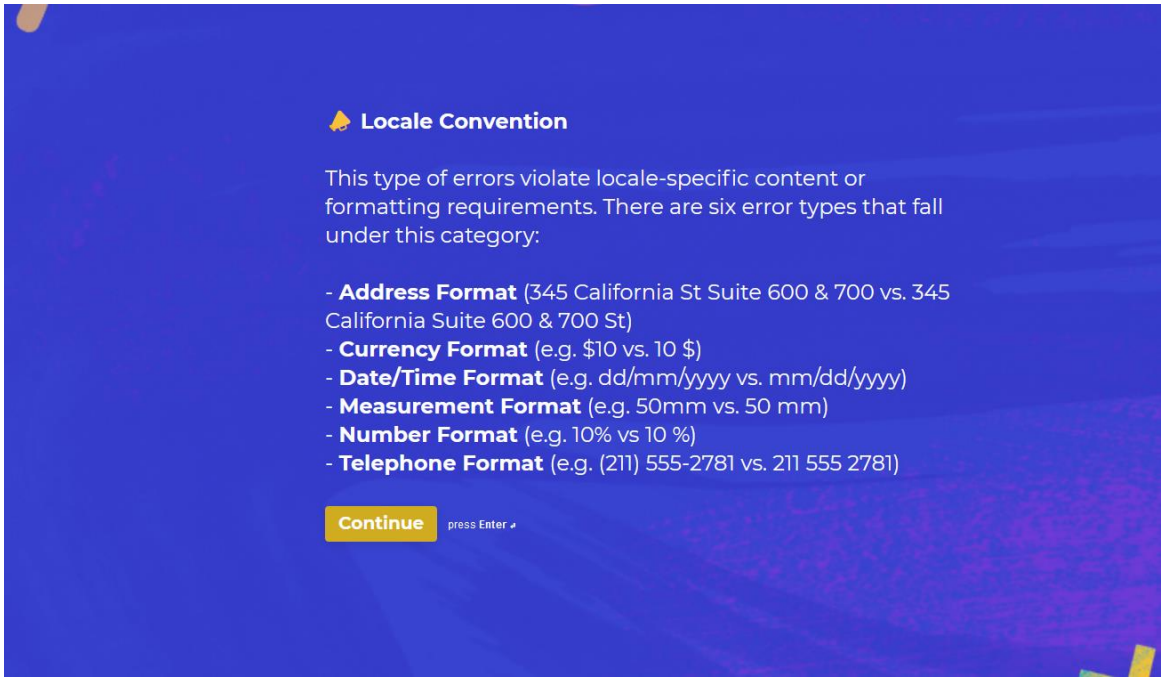

c. Would you like to see the information about Mistranslation and Unnatural Flow again to **better understand your mistake**?
Y Yes
N No ✓
OK ✓

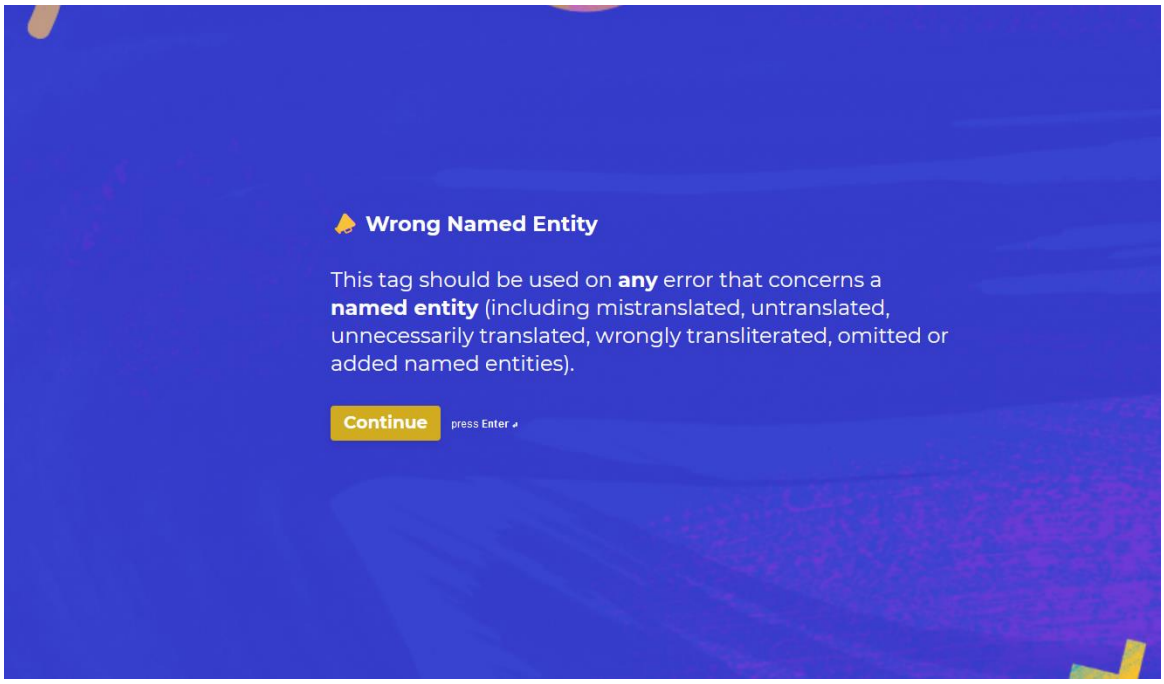

Let's now answer some questions related to Named Entities!
Continue
press Enter

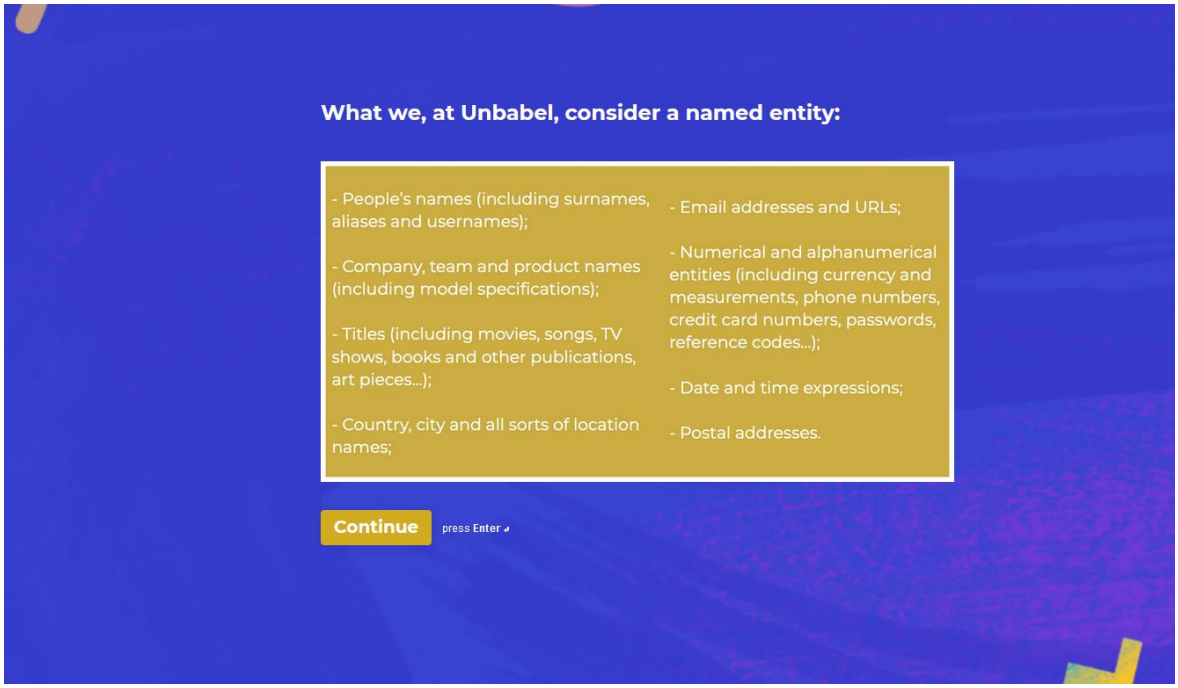

a. How would you annotate this error?
Source [EN]
Target with annotation [PT-PT]
Context
"It has a starting price of $49.99 a year."
"Tem um preço inicial de $49.99 por ano."
In Portuguese you'd translate it as "49,99 $" (dollar sign in the wrong place, period instead of comma and missing whitespace).
A Untranslated
B Wrong Named Entity
C Currency Format
OK

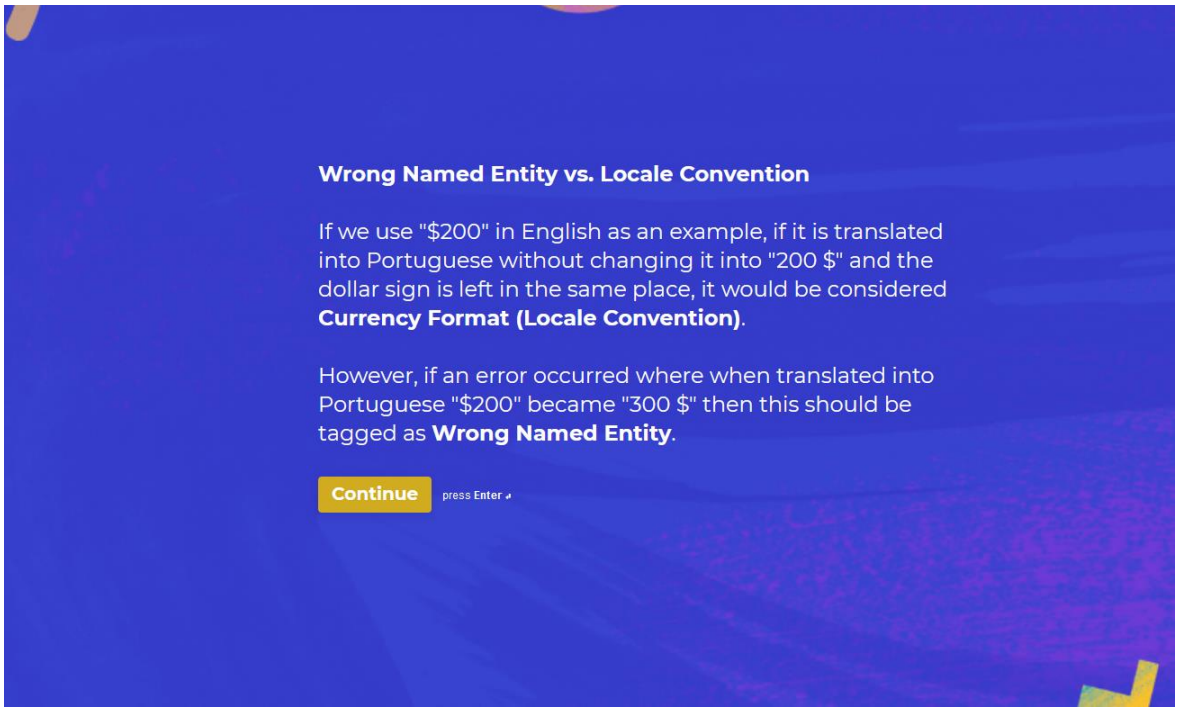

❌ Not really...
When named entities are left untranslated (i.e. a date), it may be tempting to tag this as **Untranslated**, but it would still be considered **Wrong Named Entity**.
Even so, while "$49.99" is a named entity, it should be tagged as **Currency Format**. The reason for this is that what is wrong is not the named entity itself, but the format of the currency.
You should use **Wrong Named Entity** if the content is wrong, but if it is only a matter of format, then it falls under **Locale Convention**.
Continue
press Enter

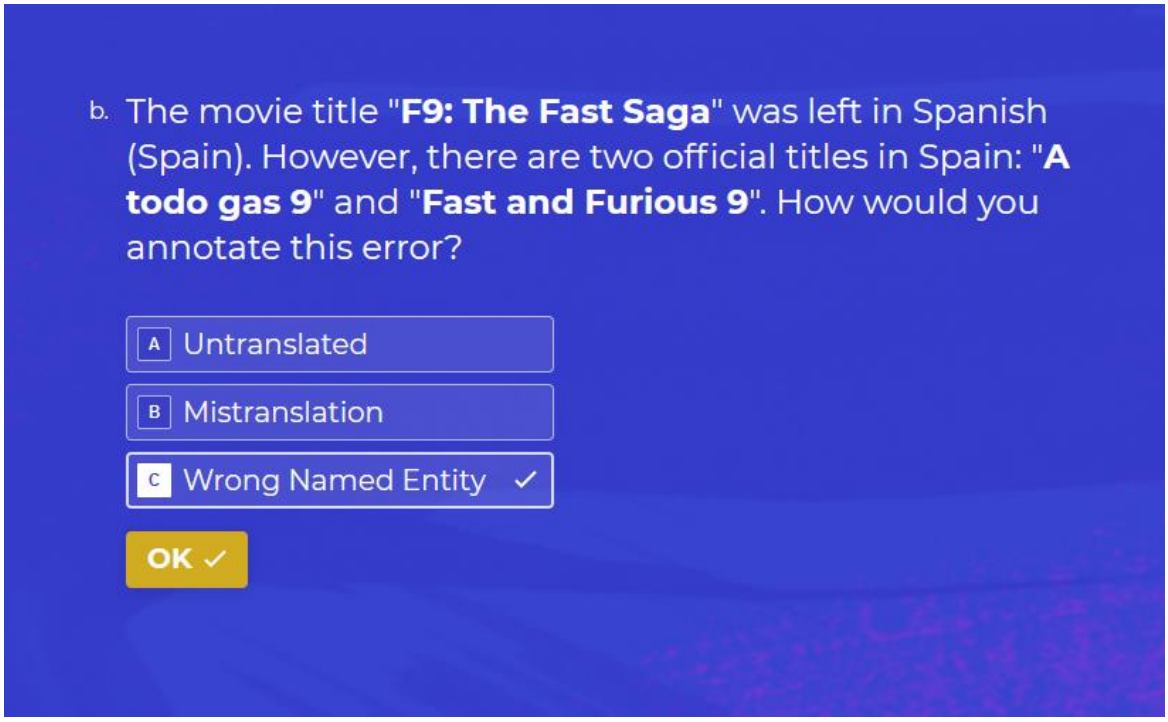

❌ Not quite!
While "$49.99" is a Named Entity, it should be tagged as **Currency Format**. The reason for this is that what is wrong is not the named entity itself, but the format of the currency.
You should use **Wrong Named Entity** if the **content** is wrong, but if it is only a matter of **format**, then it falls under **Locale Convention**.
Continue
press Enter

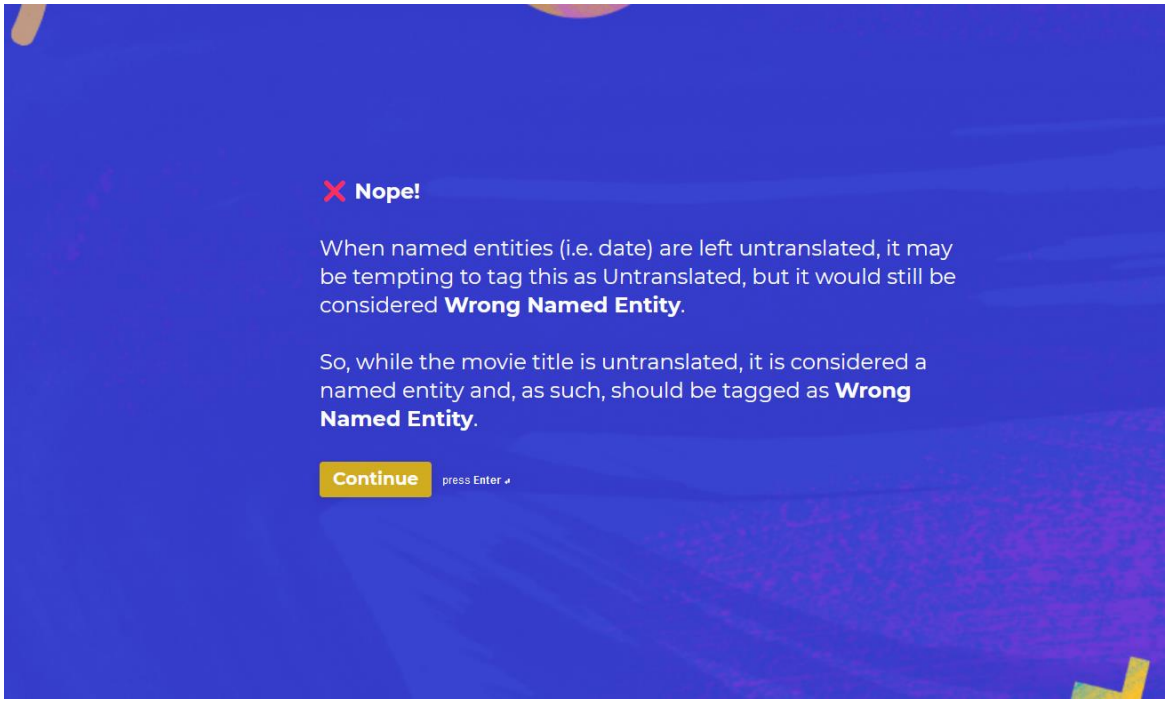

Locale Convention
This type of errors violate locale-specific content or formatting requirements. There are six error types that fall under this category:
- Address Format (345 California St Suite 600 & 700 vs. 345 California Suite 600 & 700 St)
- Currency Format (e.g. $10 vs. 10 $)
- Date/Time Format (e.g. dd/mm/yyyy vs. mm/dd/yyyy)
- Measurement Format (e.g. 50mm vs. 50 mm)
- Number Format (e.g. 10% vs 10 %)
- Telephone Format (e.g. (211) 555-2781 vs. 211 555 2781)
Continue
press Enter

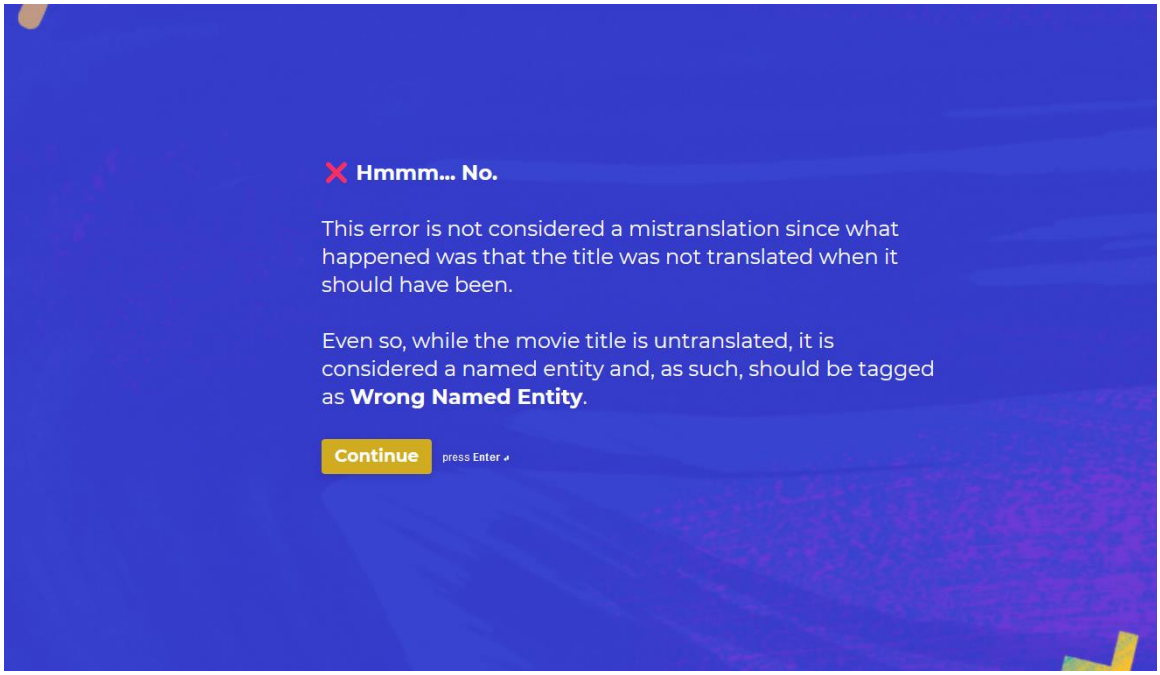

Wrong Named Entity
This tag should be used on any error that concerns a named entity (including mistranslated, untranslated, unnecessarily translated, wrongly transliterated, omitted or added named entities).
Continue
press Enter

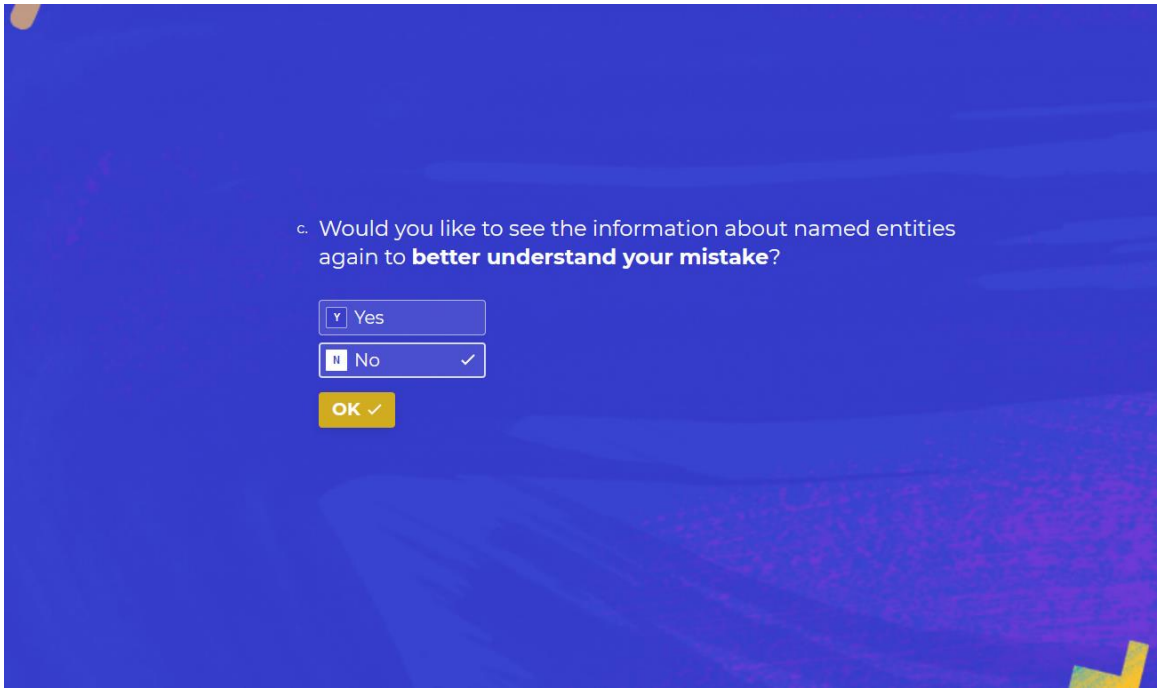

What we, at Unbabel, consider a named entity:
- People's names (including surnames, aliases and usernames);
- Company, team and product names (including model specifications);
- Titles (including movies, songs, TV shows, books and other publications, art pieces...);
- Country, city and all sorts of location names;
- Email addresses and URLs;
- Numerical and alphanumerical entities (including currency and measurements, phone numbers, credit card numbers, passwords, reference codes...);
- Date and time expressions;
- Postal addresses.
Continue
press Enter

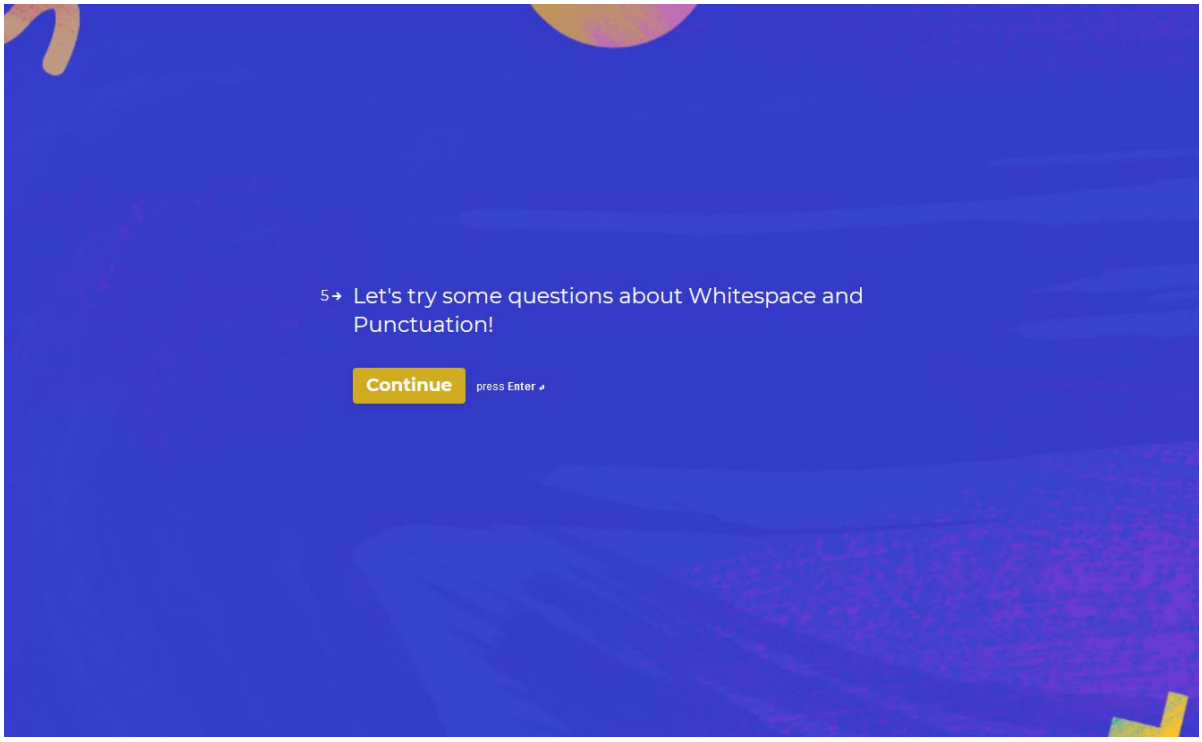

Wrong Named Entity vs. Locale Convention
If we use "$200" in English as an example, if it is translated into Portuguese without changing it into "200 $" and the dollar sign is left in the same place, it would be considered **Currency Format (Locale Convention)**.
However, if an error occurred where when translated into Portuguese "$200" became "300 $" then this should be tagged as **Wrong Named Entity**.
Continue
press Enter


b. The movie title "F9: The Fast Saga" was left in Spanish (Spain). However, there are two official titles in Spain: "A todo gas 9" and "Fast and Furious 9". How would you annotate this error?
A Untranslated
B Mistranslation
C Wrong Named Entity ✓
OK ✓


✗ Nope!
When named entities (i.e. date) are left untranslated, it may be tempting to tag this as Untranslated, but it would still be considered Wrong Named Entity.
So, while the movie title is untranslated, it is considered a named entity and, as such, should be tagged as Wrong Named Entity.
Continue
press Enter


Hmmm... No.
This error is not considered a mistranslation since what happened was that the title was not translated when it should have been.
Even so, while the movie title is untranslated, it is considered a named entity and, as such, should be tagged as **Wrong Named Entity**.
Continue
press Enter


c. Would you like to see the information about named entities again to **better understand your mistake**?
Y Yes
N No
OK



a. How would you annotate this error?

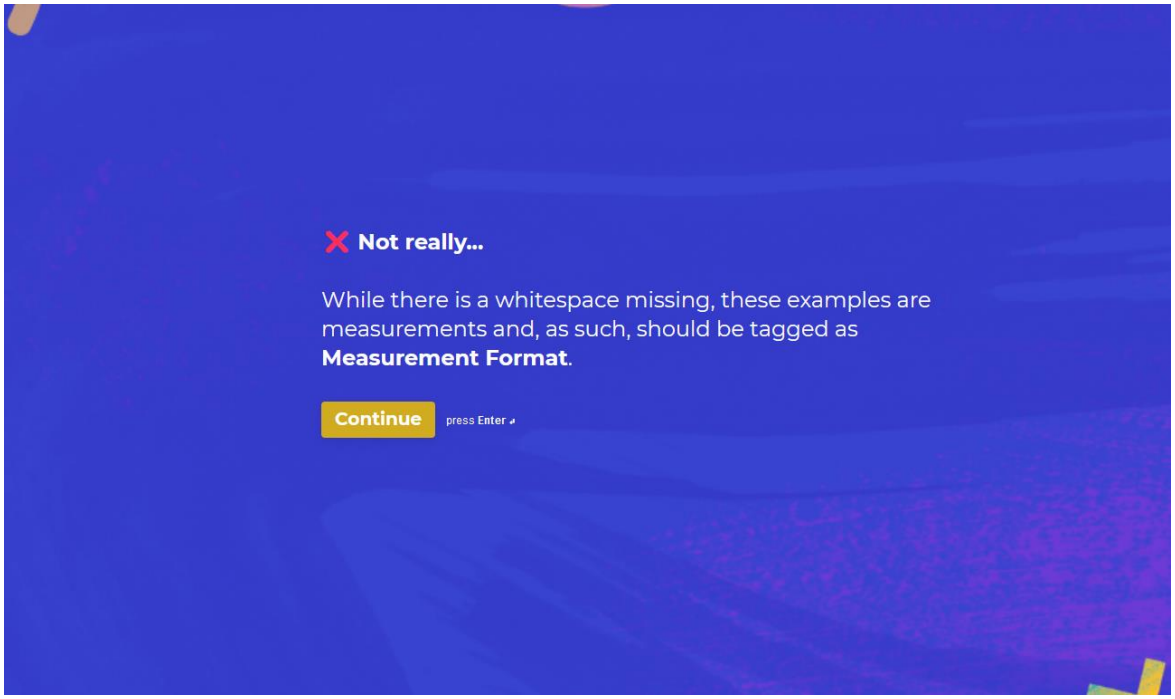
Source [EN]	Target with annotation [SV]	Context
"A 4.8V threshold voltage that allows 0V to 15V drive."	"På 4.8V som möjliggör drivenhet mellan 0V och 15V."	In Swedish, there should be a space between the number and the 'V'.

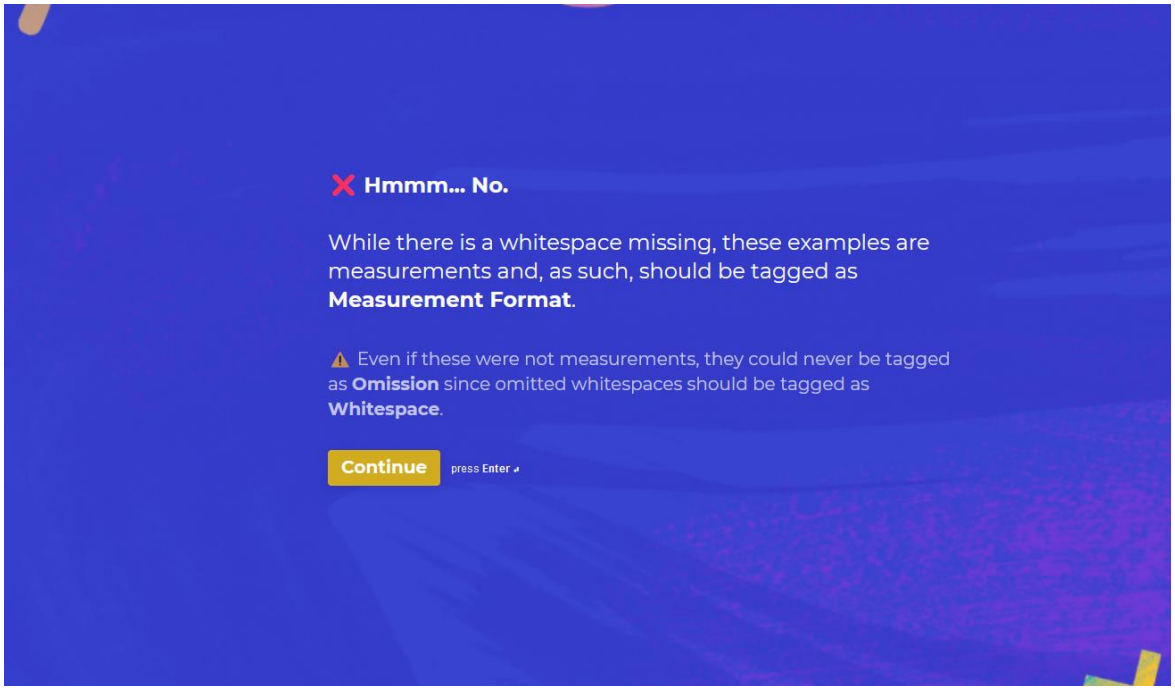
A Whitespace

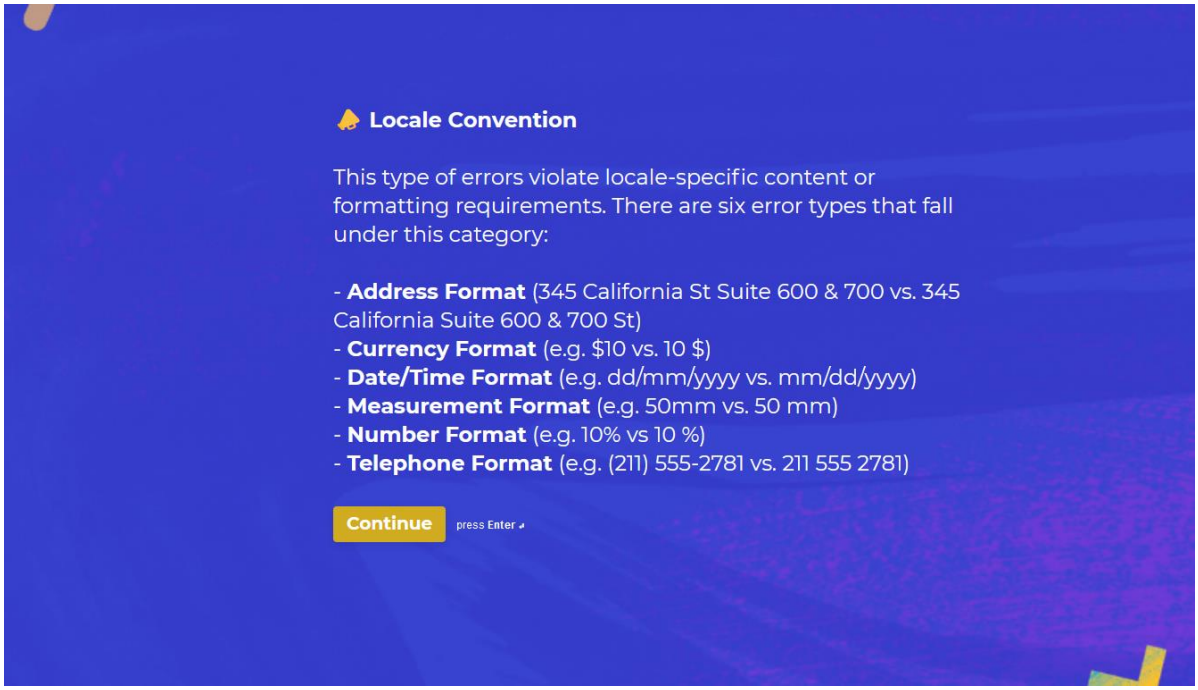
B Omission

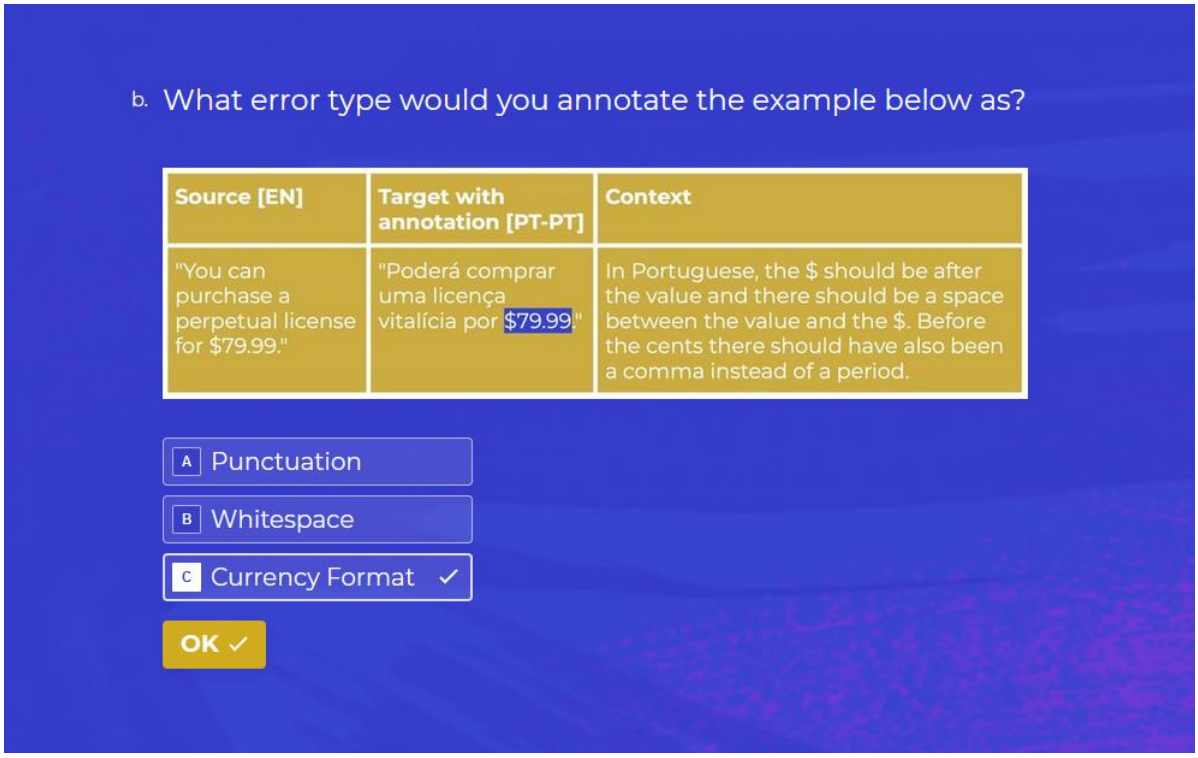
C Measurement Format ✓

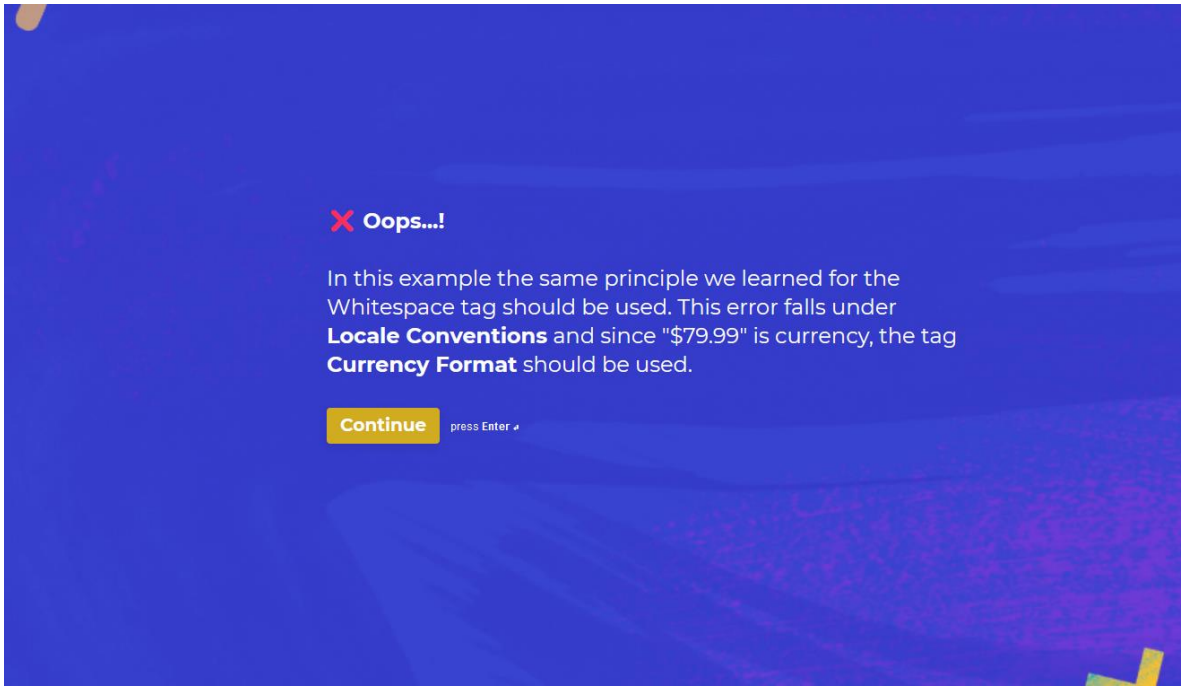
OK ✓

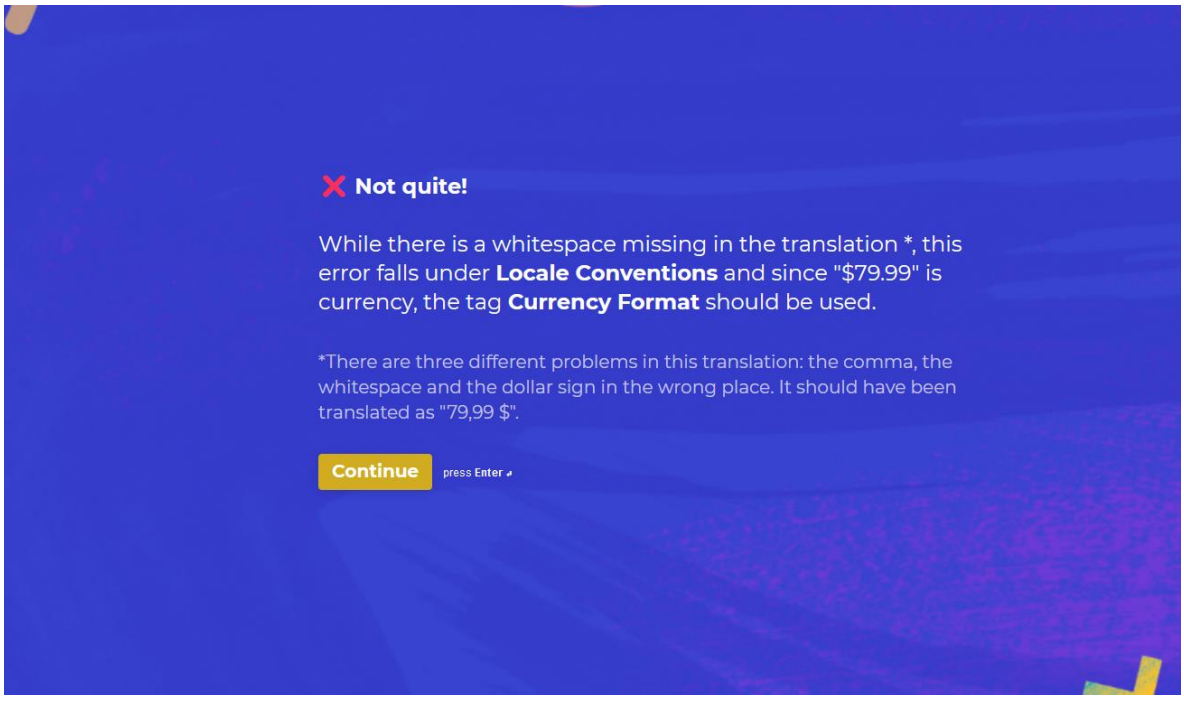

❌ Not really...
While there is a whitespace missing, these examples are measurements and, as such, should be tagged as **Measurement Format**.
Continue
press Enter

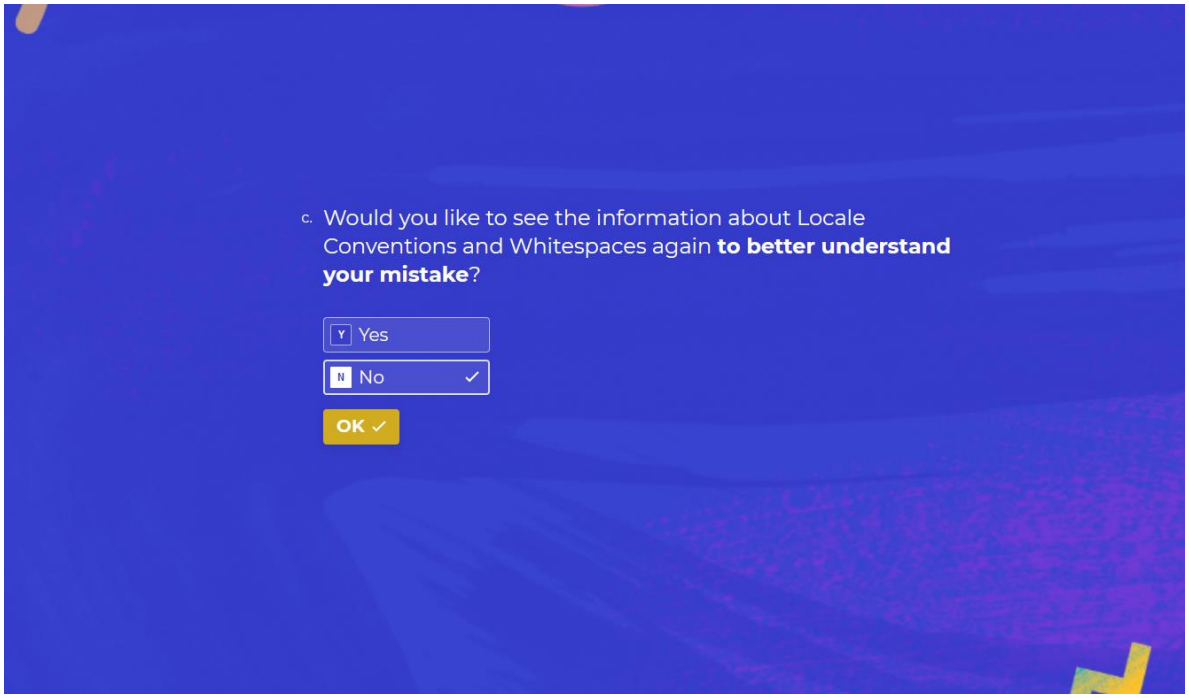

❌ Hmmm... No.
While there is a whitespace missing, these examples are measurements and, as such, should be tagged as **Measurement Format**.
⚠ Even if these were not measurements, they could never be tagged as **Omission** since omitted whitespaces should be tagged as **Whitespace**.
Continue
press Enter

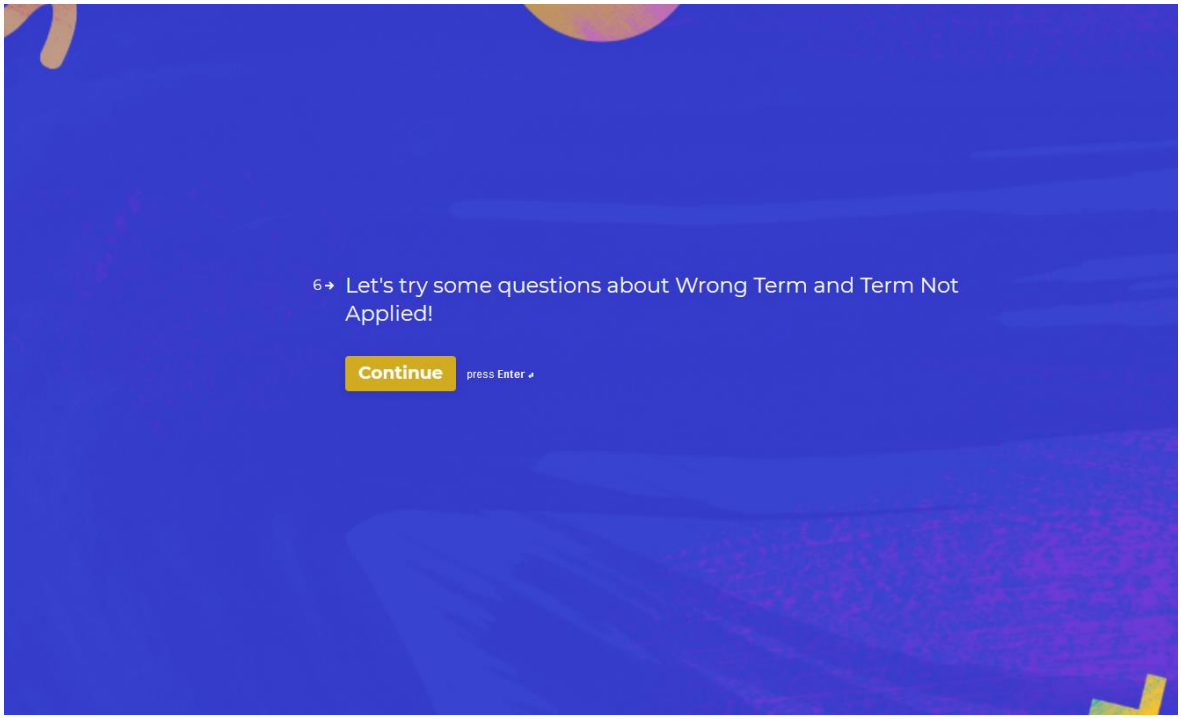

Locale Convention
This type of errors violate locale-specific content or formatting requirements. There are six error types that fall under this category:
- Address Format (345 California St Suite 600 & 700 vs. 345 California Suite 600 & 700 St)
- Currency Format (e.g. $10 vs. 10 $)
- Date/Time Format (e.g. dd/mm/yyyy vs. mm/dd/yyyy)
- Measurement Format (e.g. 50mm vs. 50 mm)
- Number Format (e.g. 10% vs 10 %)
- Telephone Format (e.g. (211) 555-2781 vs. 211 555 2781)
Continue
press Enter


b. What error type would you annotate the example below as?
Source [EN]
Target with annotation [PT-PT]
Context
"You can purchase a perpetual license for $79.99."
"Poderá comprar uma licença vitalícia por $79.99"
In Portuguese, the $ should be after the value and there should be a space between the value and the $. Before the cents there should have also been a comma instead of a period.
A Punctuation
B Whitespace
C Currency Format
OK


Oops...!
In this example the same principle we learned for the Whitespace tag should be used. This error falls under **Locale Conventions** and since "$79.99" is currency, the tag **Currency Format** should be used.
Continue
press Enter


Not quite!
While there is a whitespace missing in the translation *, this error falls under **Locale Conventions** and since "$79.99" is currency, the tag **Currency Format** should be used.
*There are three different problems in this translation: the comma, the whitespace and the dollar sign in the wrong place. It should have been translated as "79,99 $".
Continue
press Enter


c. Would you like to see the information about Locale Conventions and Whitespaces again **to better understand your mistake**?
Y Yes
N No
OK


6→ Let's try some questions about Wrong Term and Term Not Applied!
Continue
press Enter

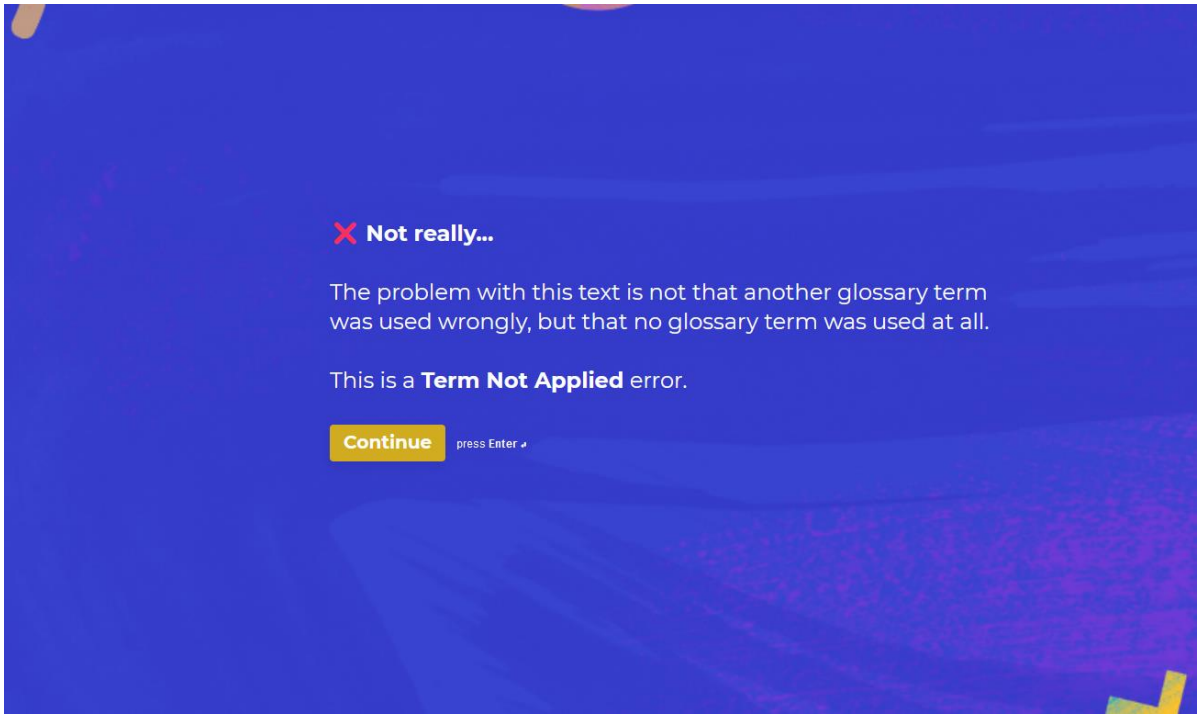
a. How would you annotate this error?

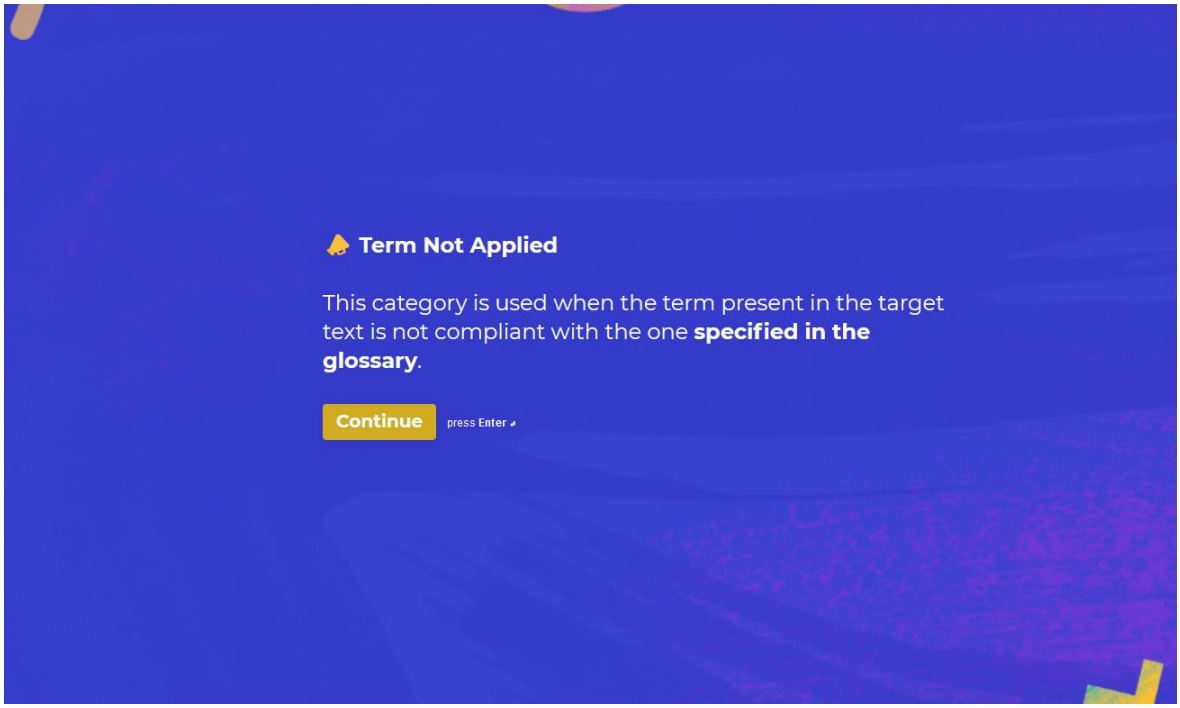
Source [EN]	Target with annotation [ES-LATAM]	Glossary
"The ROHM Semiconductor luminous intensity can range from 200mcd to 27000mcd depending on the selected emitting color."	"La intensidad luminosa de los Semiconductores ROHM puede variar de 200mcd a 27000mcd según el color de emisión seleccionado."	Source term: ROHM Semiconductor Expected translation: ROHM Semiconductor

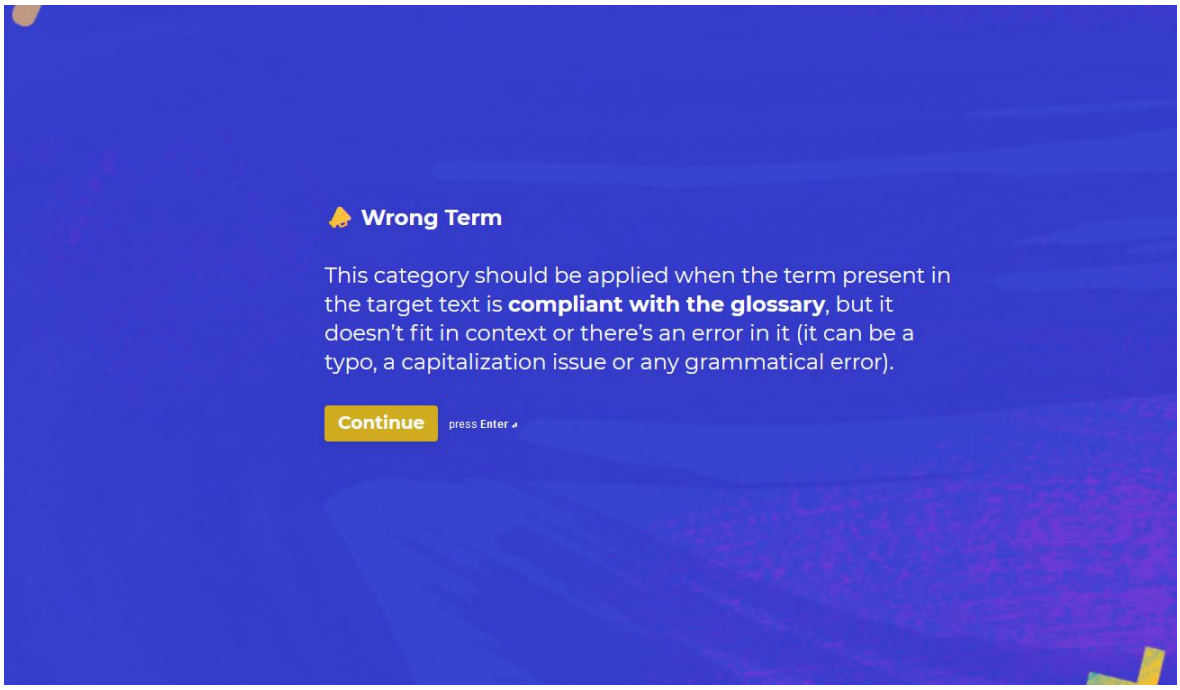
A Wrong Term

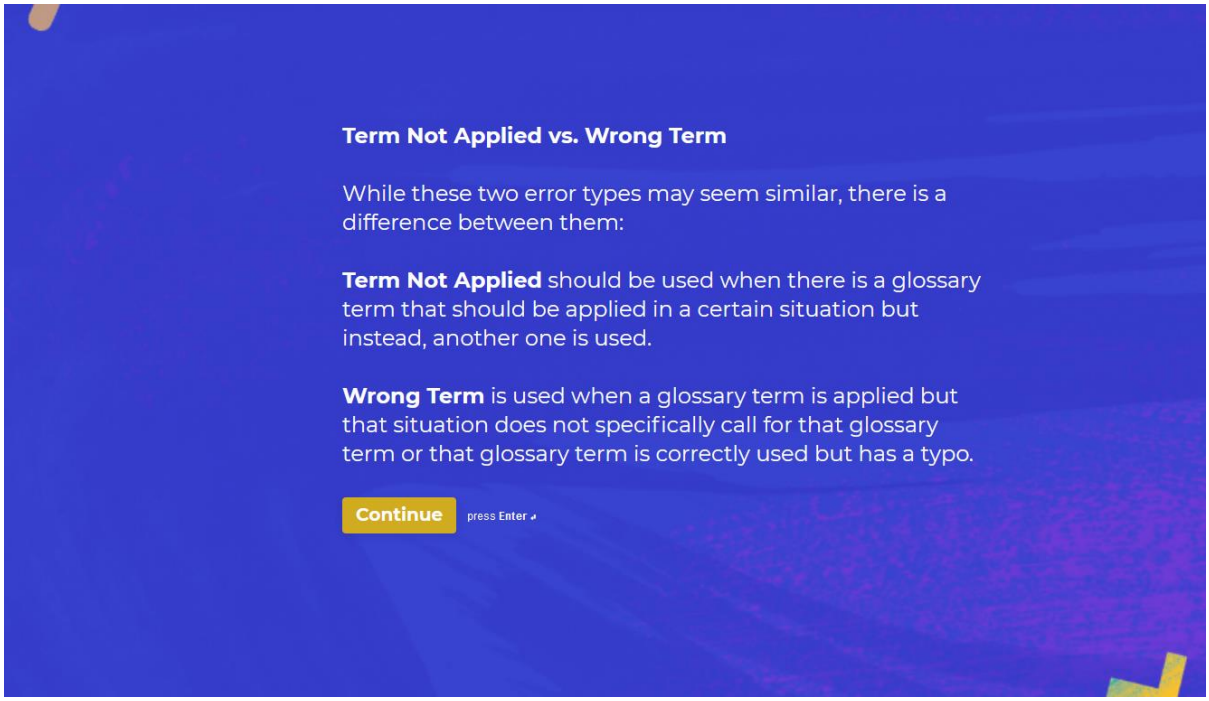
B Term Not Applied ✓

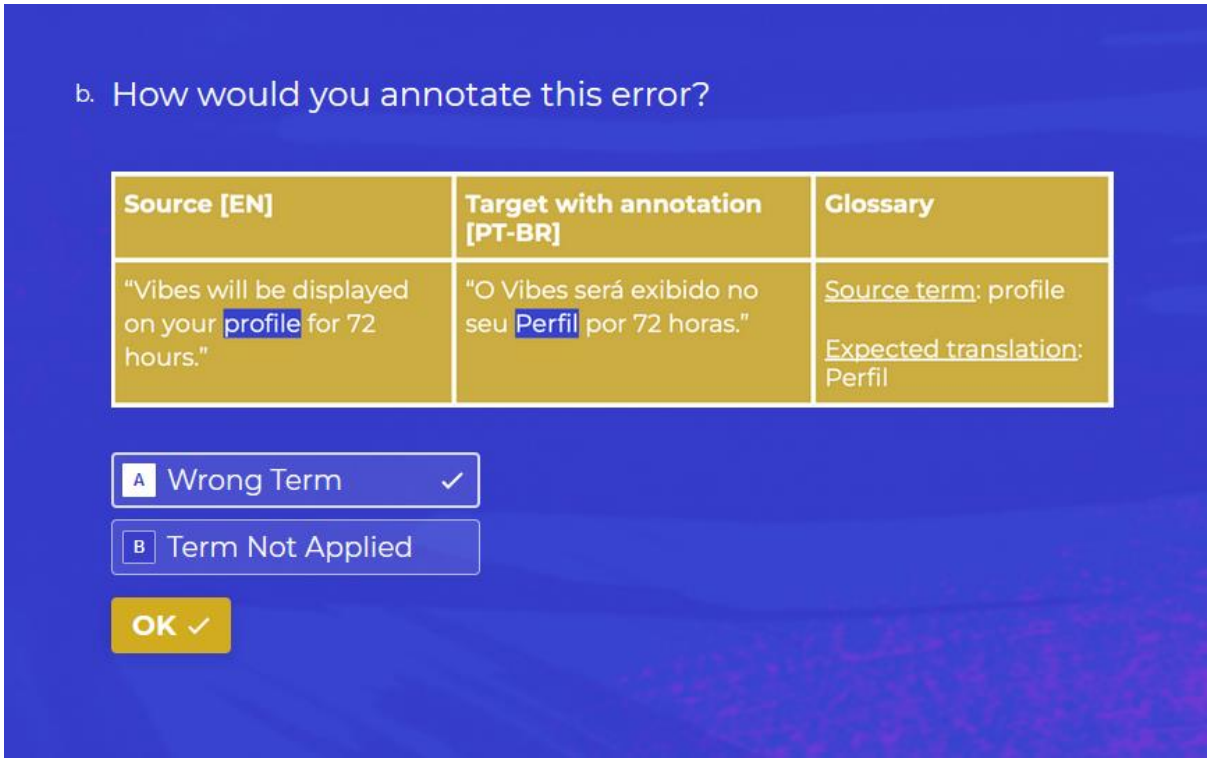
OK ✓

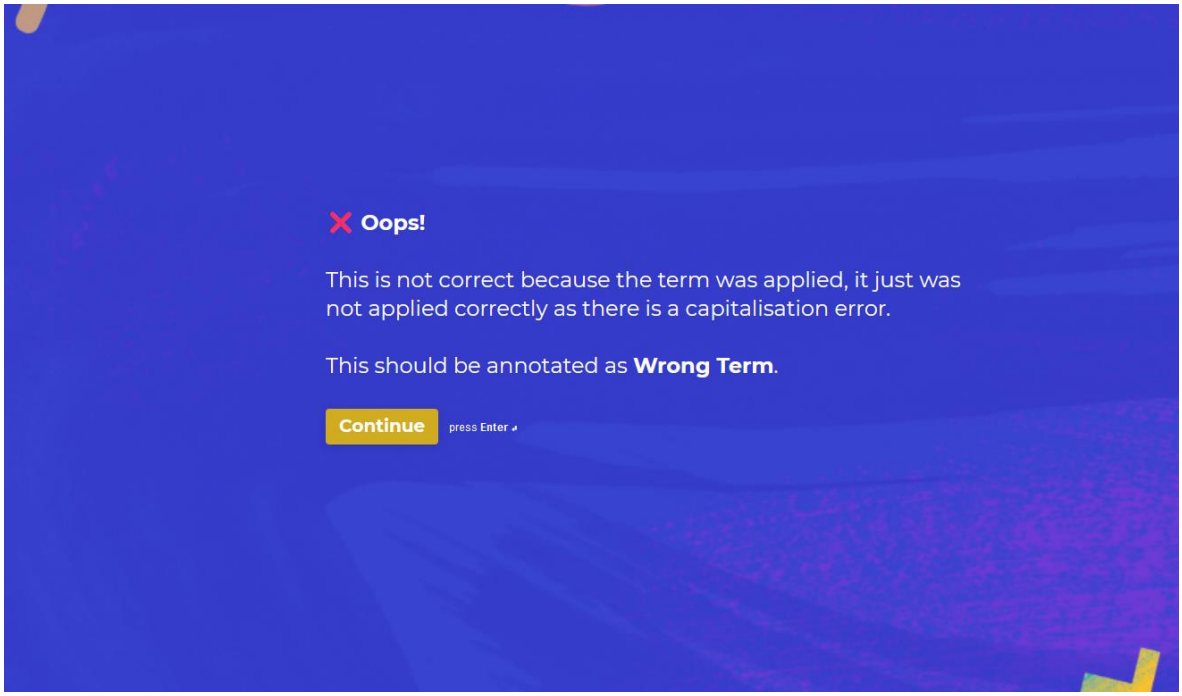


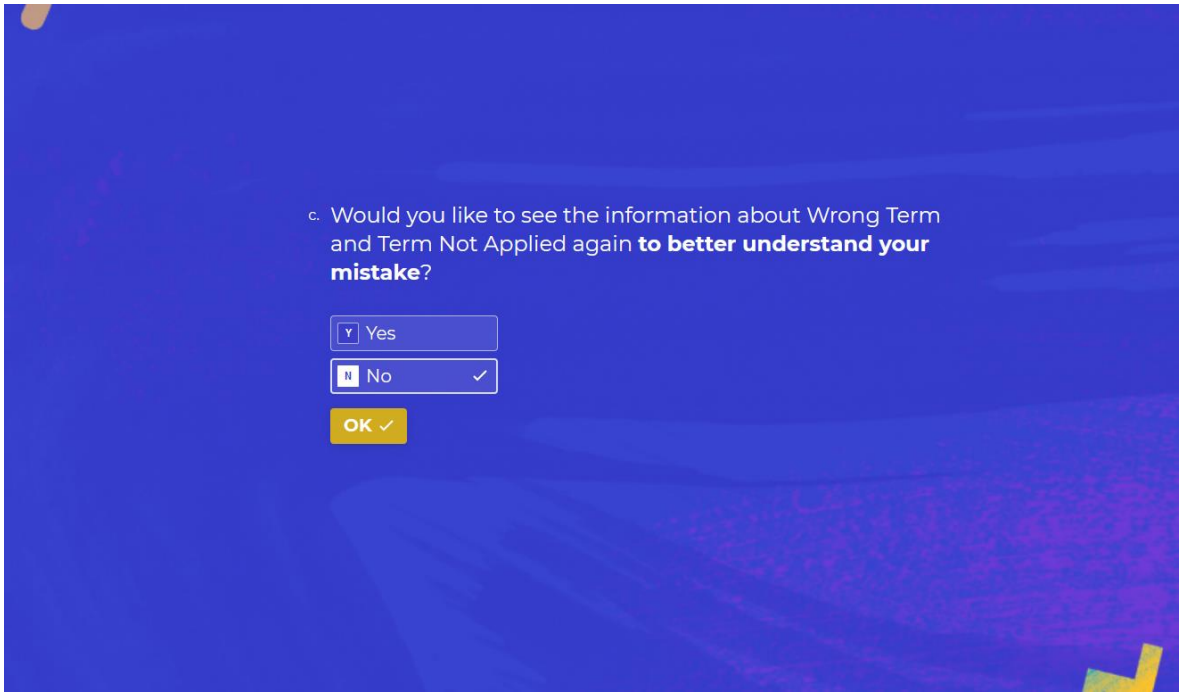

Term Not Applied
This category is used when the term present in the target text is not compliant with the one specified in the glossary.
Continue
press Enter

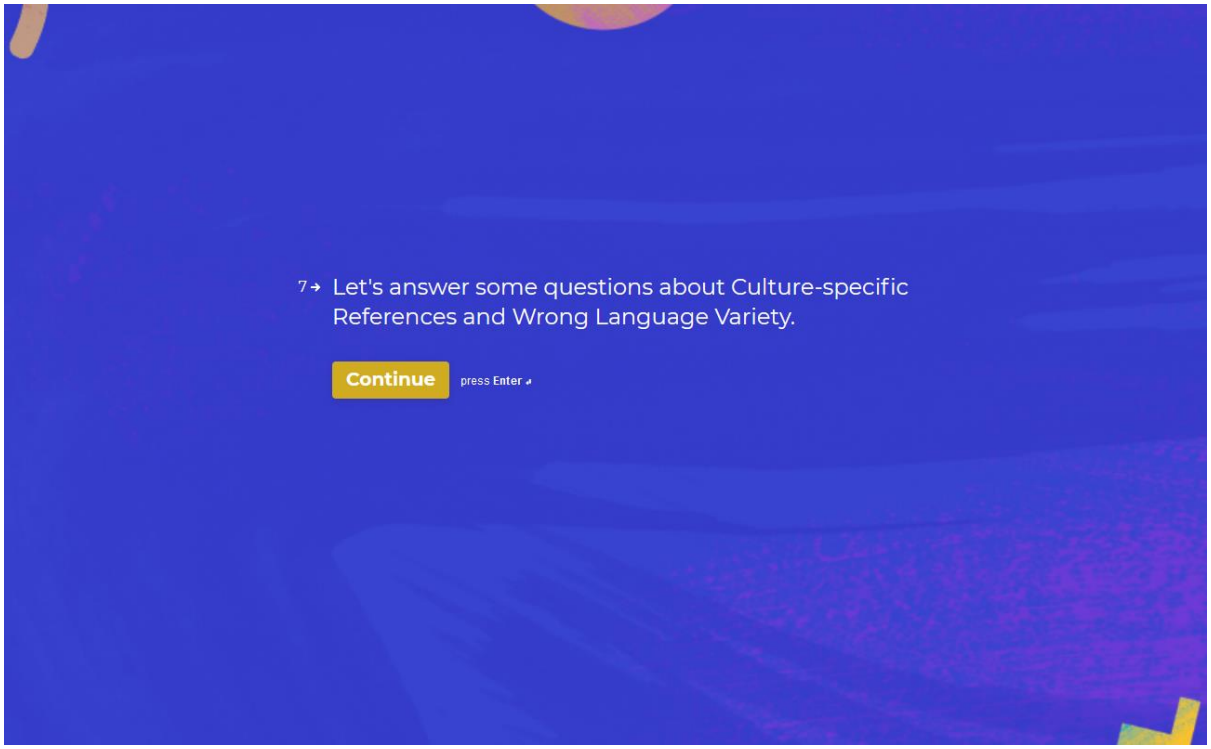

Wrong Term
This category should be applied when the term present in the target text is compliant with the glossary, but it doesn't fit in context or there's an error in it (it can be a typo, a capitalization issue or any grammatical error).
Continue
press Enter

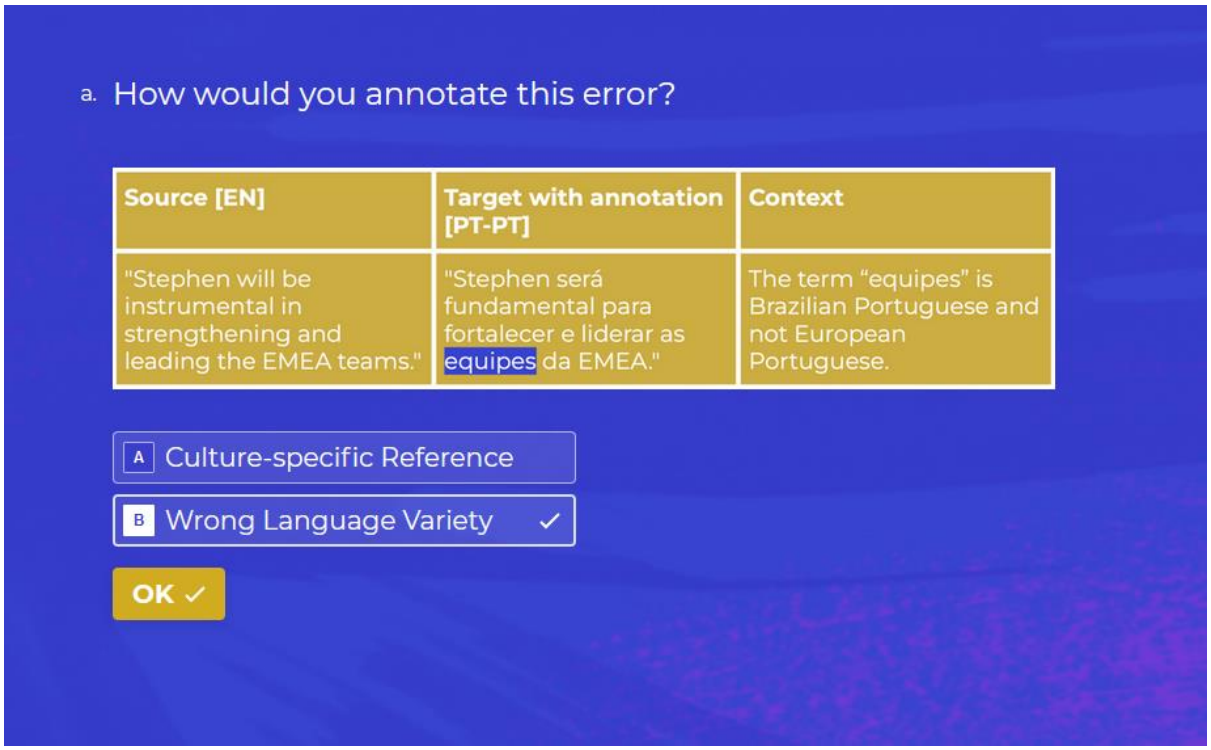

Term Not Applied vs. Wrong Term
While these two error types may seem similar, there is a difference between them:
Term Not Applied should be used when there is a glossary term that should be applied in a certain situation but instead, another one is used.
Wrong Term is used when a glossary term is applied but that situation does not specifically call for that glossary term or that glossary term is correctly used but has a typo.
Continue
press Enter

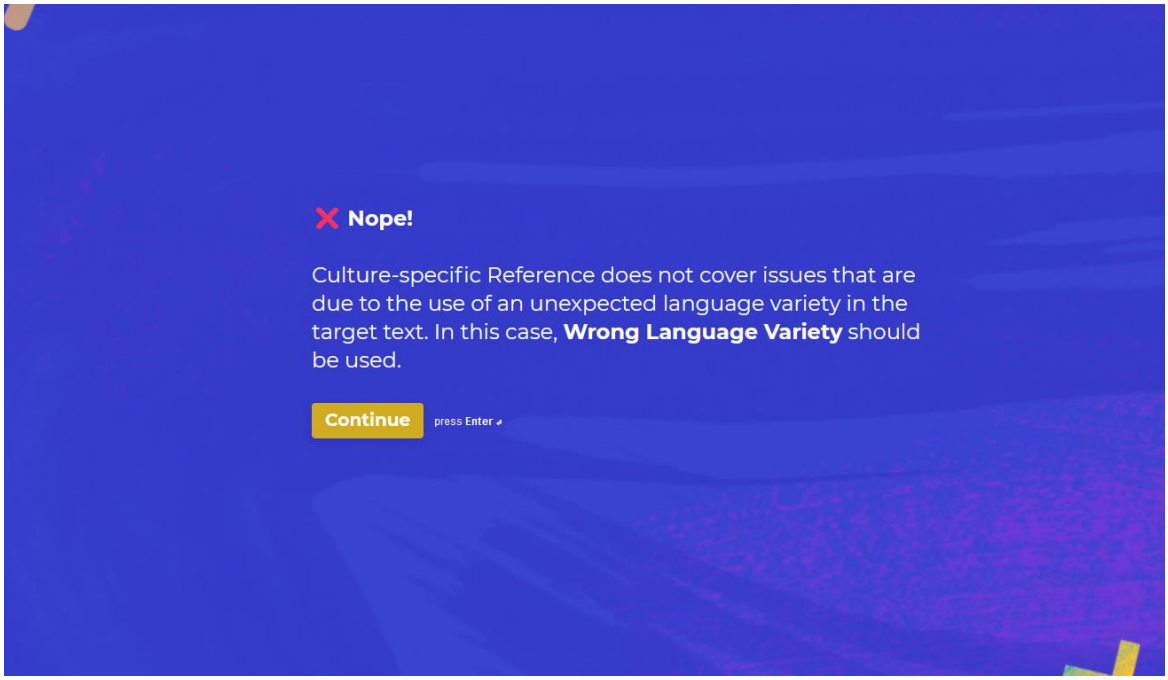

b. How would you annotate this error?
Source [EN]
Target with annotation [PT-BR]
Glossary
"Vibes will be displayed on your profile for 72 hours."
"O Vibes será exibido no seu Perfil por 72 horas."
Source term: profile
Expected translation: Perfil
A Wrong Term
B Term Not Applied
OK

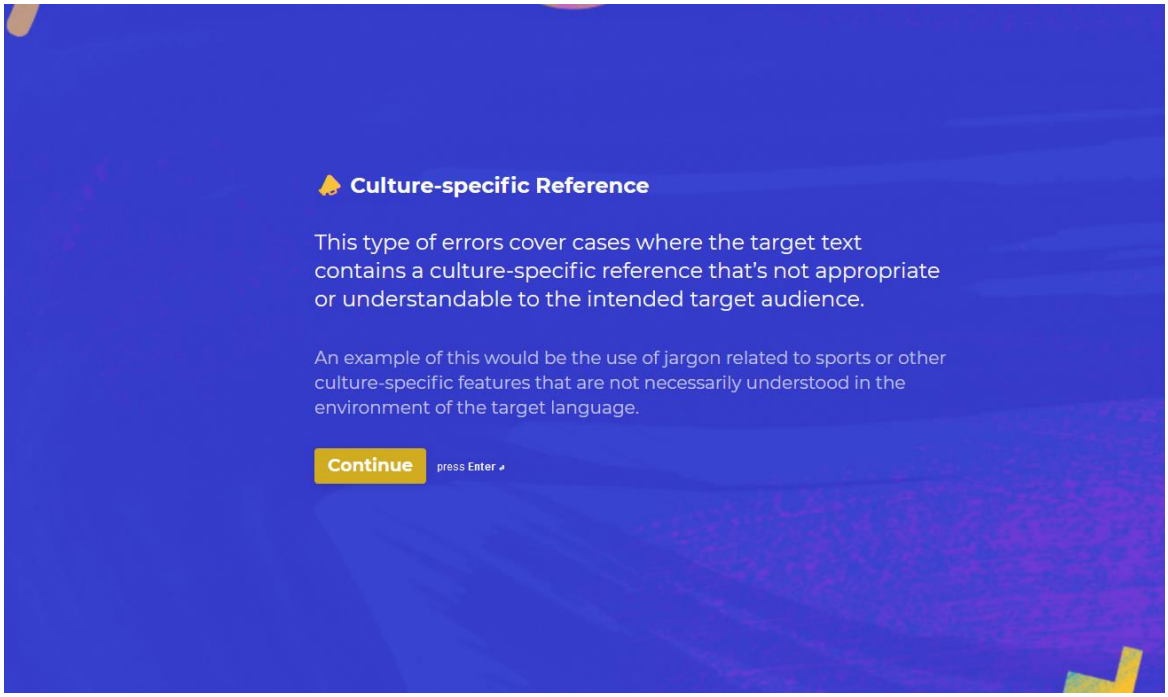

Oops!
This is not correct because the term was applied, it just was not applied correctly as there is a capitalisation error.
This should be annotated as **Wrong Term**.
Continue
press Enter

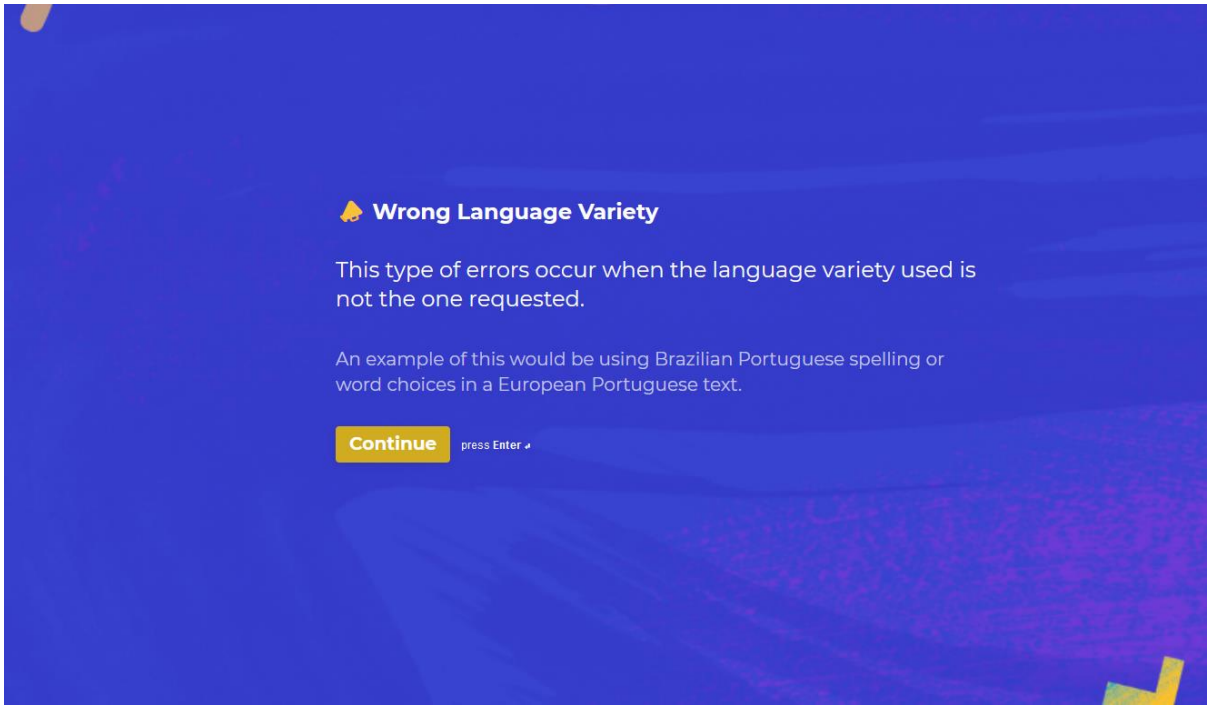

c. Would you like to see the information about Wrong Term and Term Not Applied again **to better understand your mistake**?
Y Yes
N No
OK

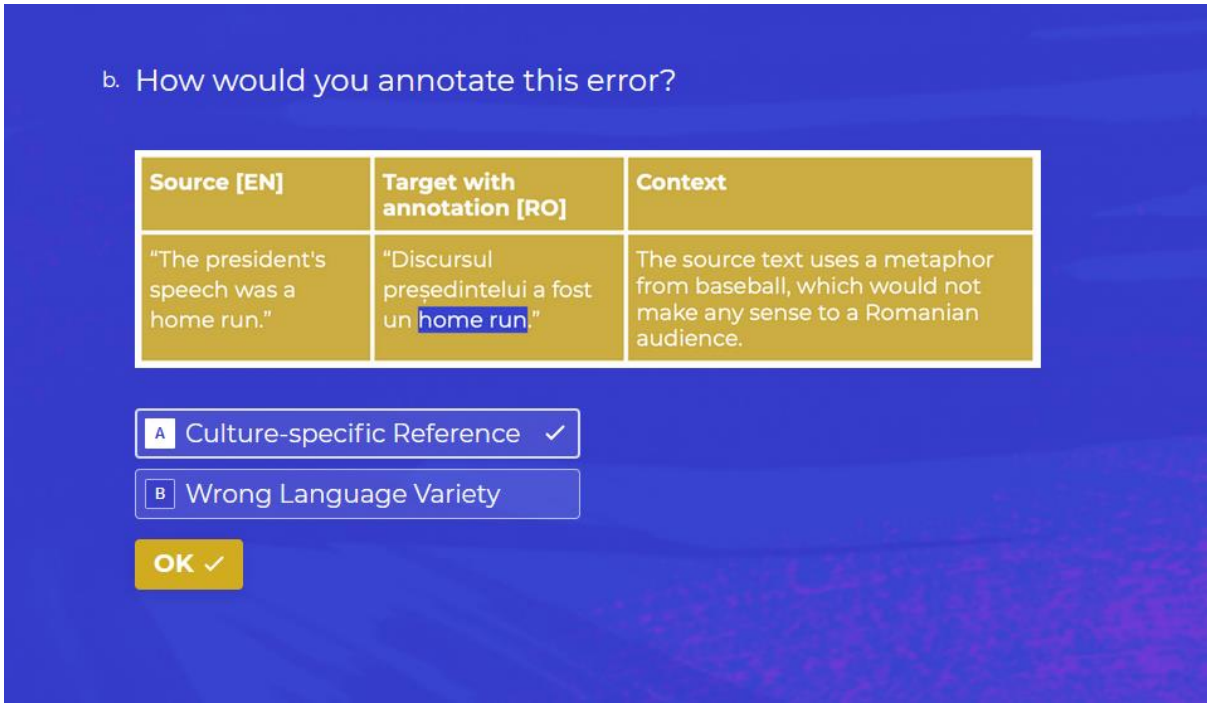

7→ Let's answer some questions about Culture-specific References and Wrong Language Variety.
Continue
press Enter

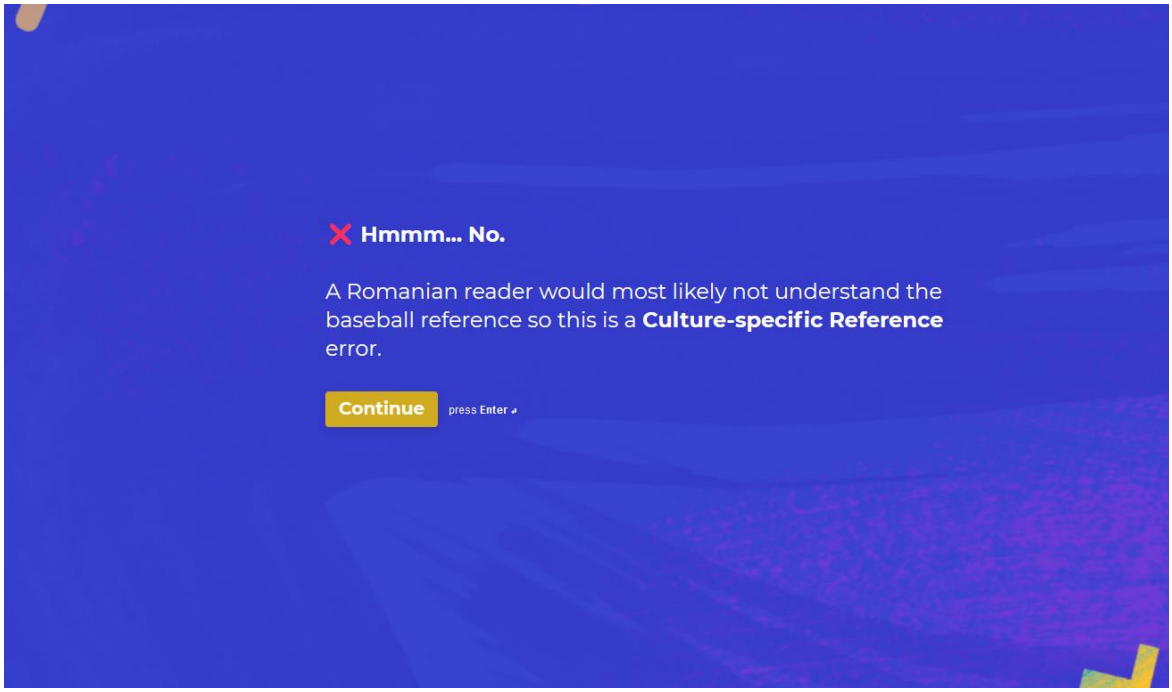

a. How would you annotate this error?
Source [EN]
Target with annotation [PT-PT]
Context
"Stephen will be instrumental in strengthening and leading the EMEA teams."
"Stephen será fundamental para fortalecer e liderar as equipes da EMEA."
The term "equipes" is Brazilian Portuguese and not European Portuguese.
A Culture-specific Reference
B Wrong Language Variety ✓
OK ✓

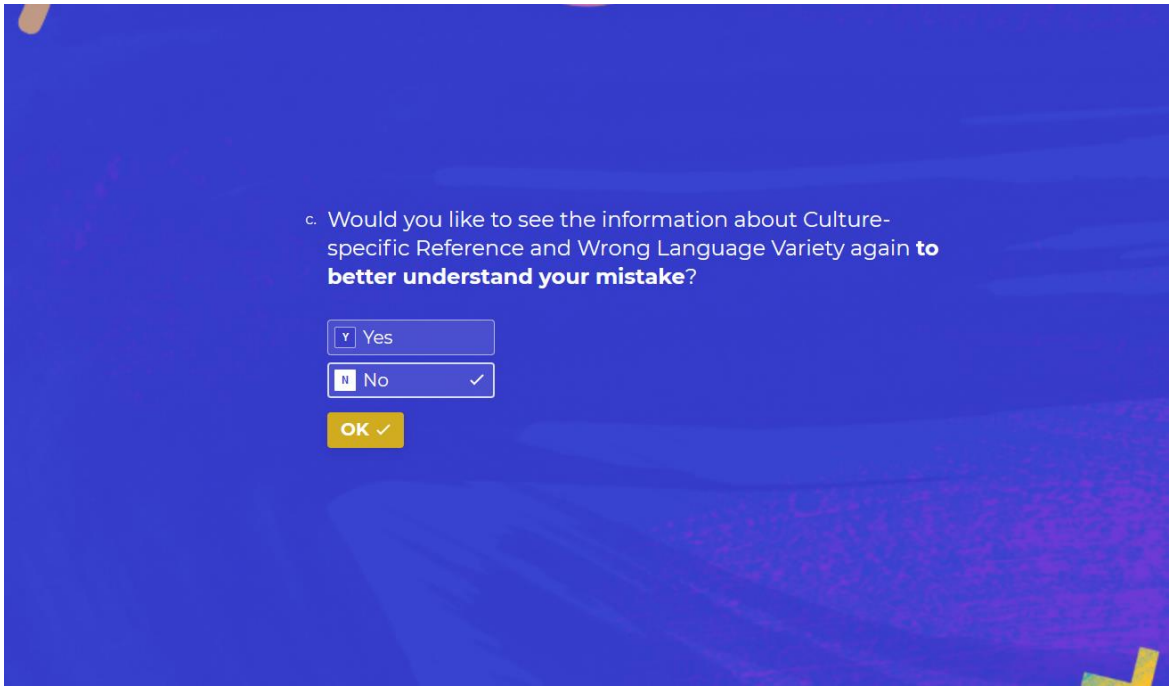

❌ Nope!
Culture-specific Reference does not cover issues that are due to the use of an unexpected language variety in the target text. In this case, **Wrong Language Variety** should be used.
Continue
press Enter

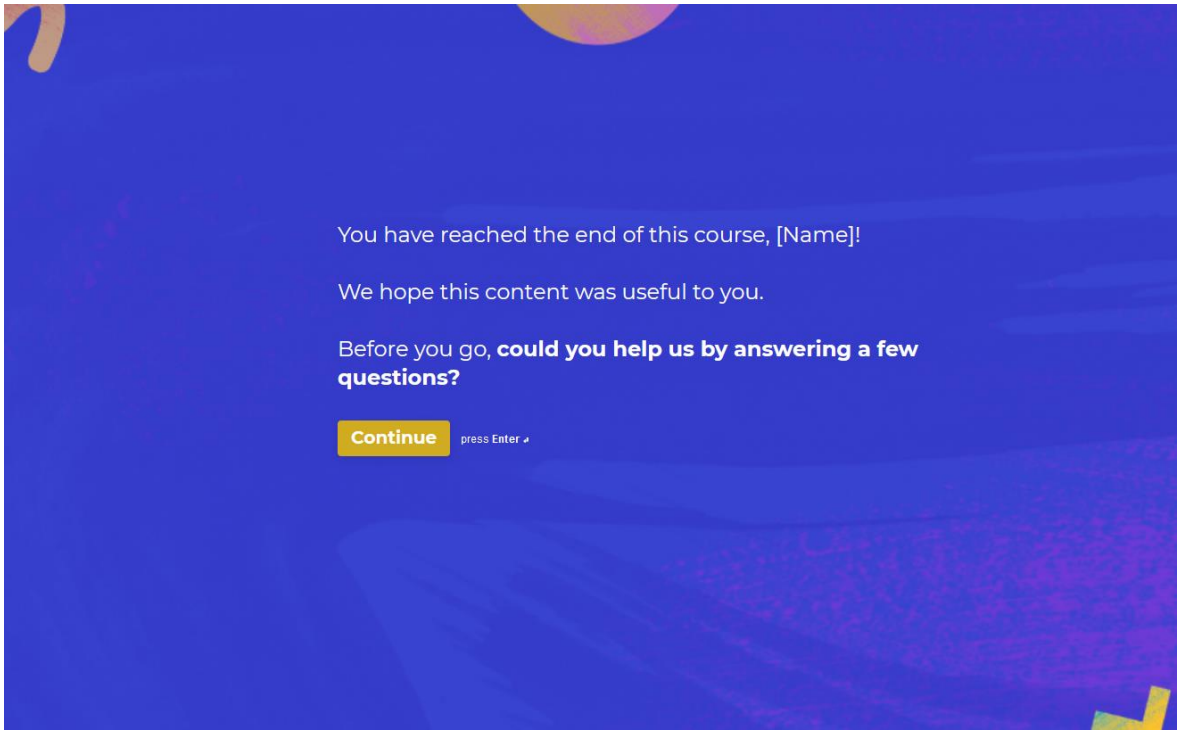

📣 Culture-specific Reference
This type of errors cover cases where the target text contains a culture-specific reference that's not appropriate or understandable to the intended target audience.
An example of this would be the use of jargon related to sports or other culture-specific features that are not necessarily understood in the environment of the target language.
Continue
press Enter

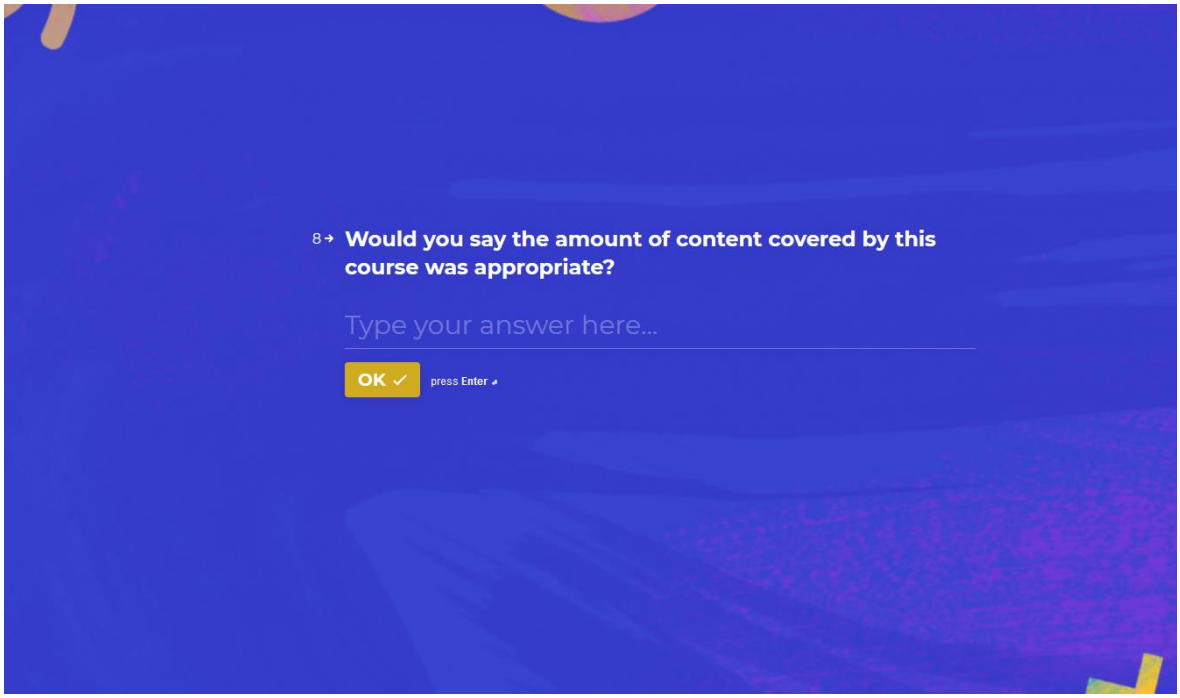

Wrong Language Variety
This type of errors occur when the language variety used is not the one requested.
An example of this would be using Brazilian Portuguese spelling or word choices in a European Portuguese text.
Continue
press Enter

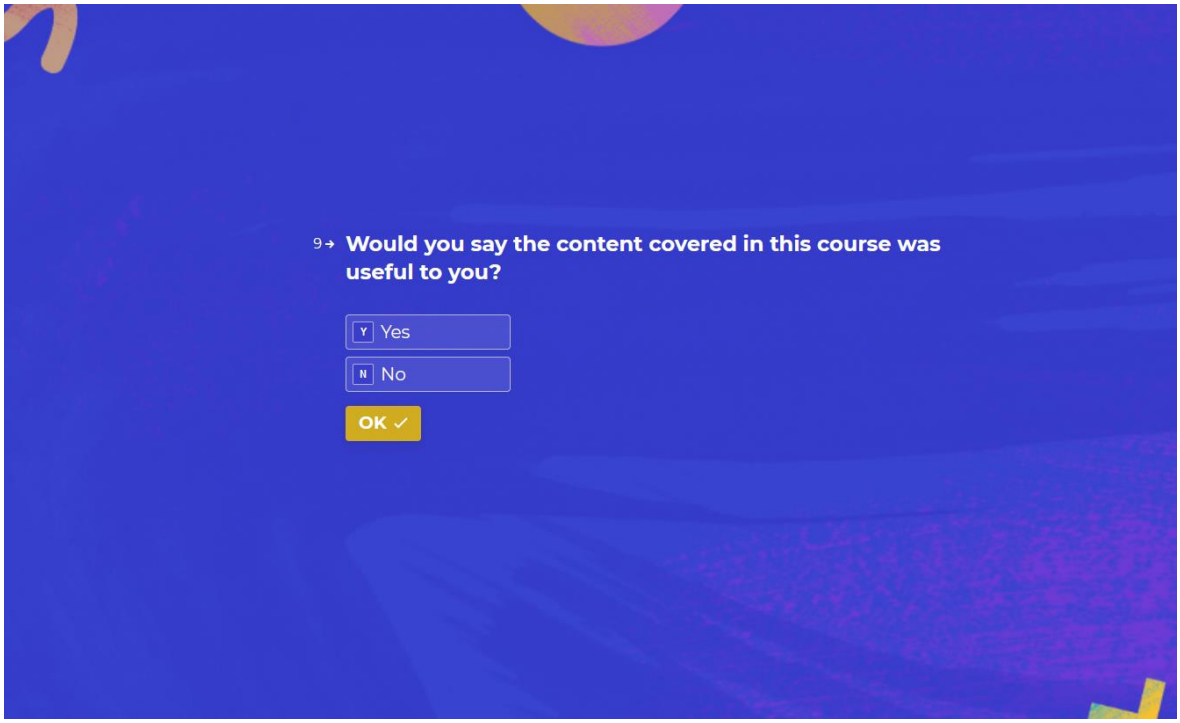

b. How would you annotate this error?
Source [EN]
Target with annotation [RO]
Context
"The president's speech was a home run."
"Discursul președintelui a fost un home run."
The source text uses a metaphor from baseball, which would not make any sense to a Romanian audience.
A Culture-specific Reference
B Wrong Language Variety
OK

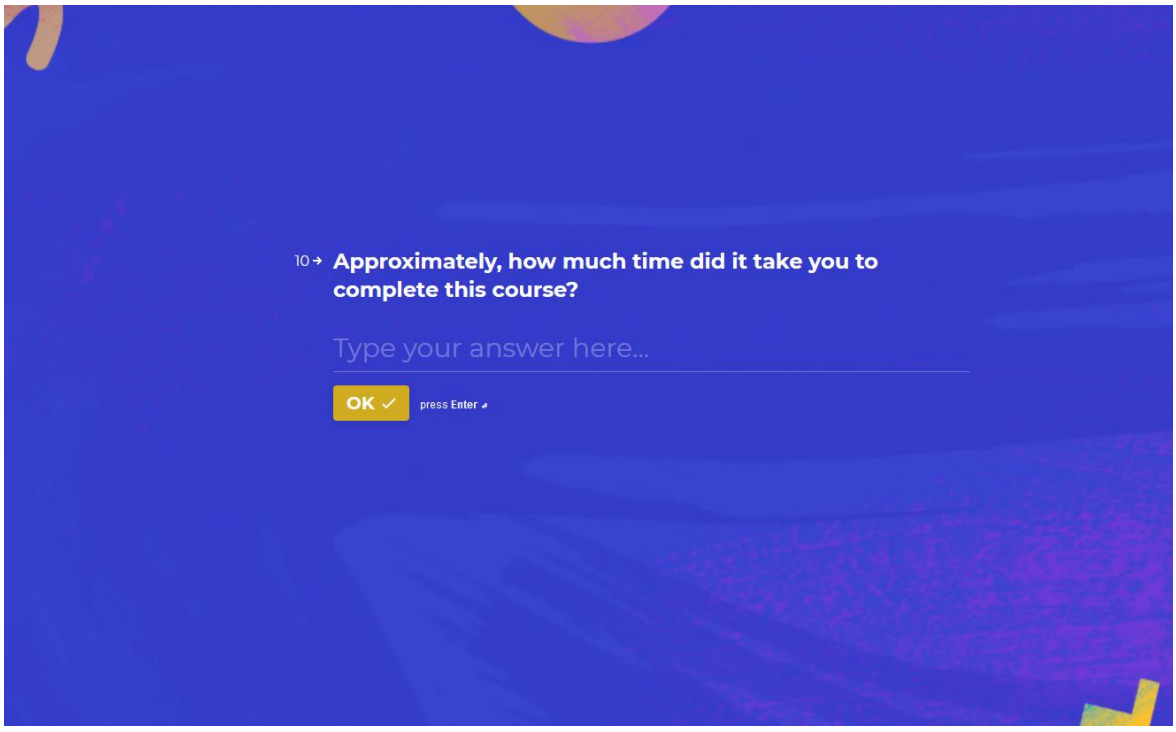

Hmmm... No.
A Romanian reader would most likely not understand the baseball reference so this is a **Culture-specific Reference** error.
Continue
press Enter

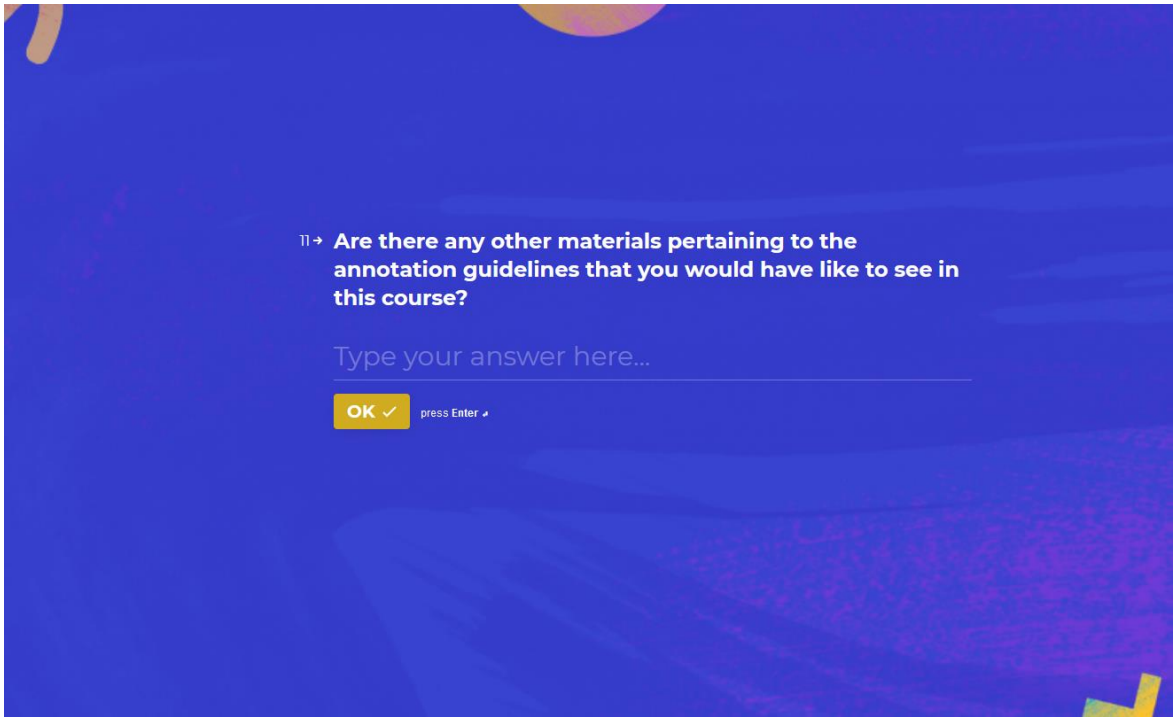

c. Would you like to see the information about Culture-specific Reference and Wrong Language Variety again **to better understand your mistake**?
Y Yes
N No
OK

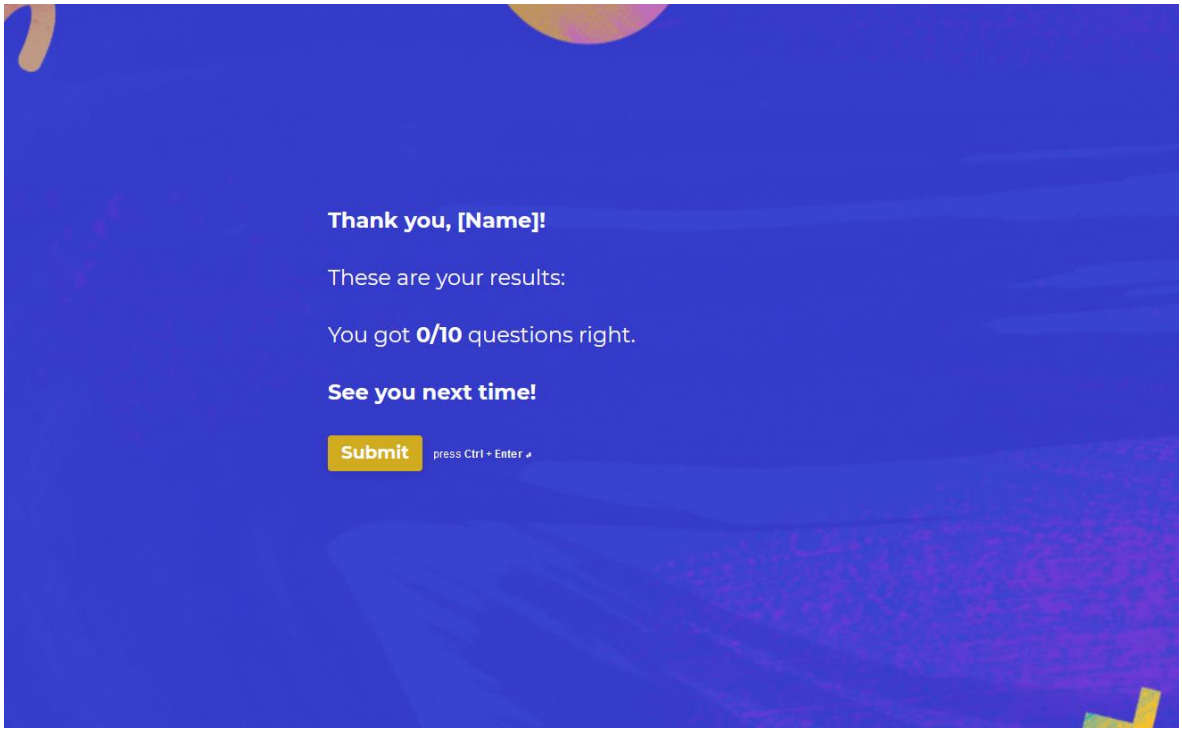

You have reached the end of this course, [Name]!
We hope this content was useful to you.
Before you go, could you help us by answering a few questions?
Continue
press Enter


8→ Would you say the amount of content covered by this course was appropriate?
Type your answer here...
OK ✓
press Enter

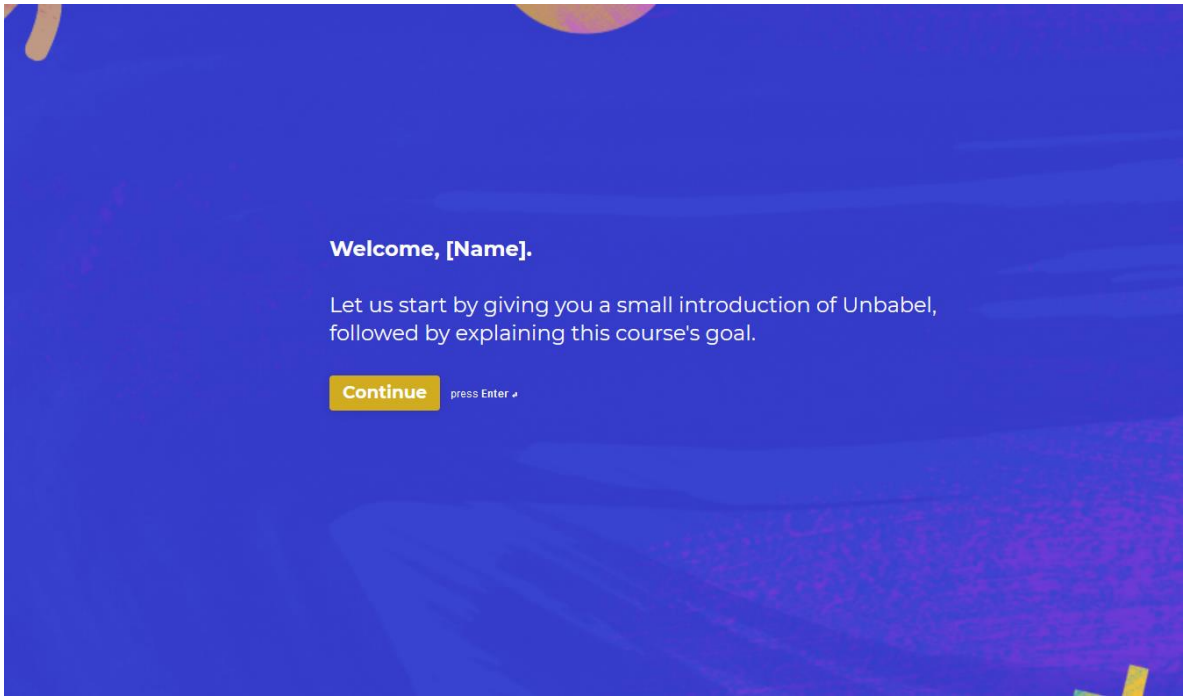

9→ Would you say the content covered in this course was useful to you?
Y Yes
N No
OK ✓

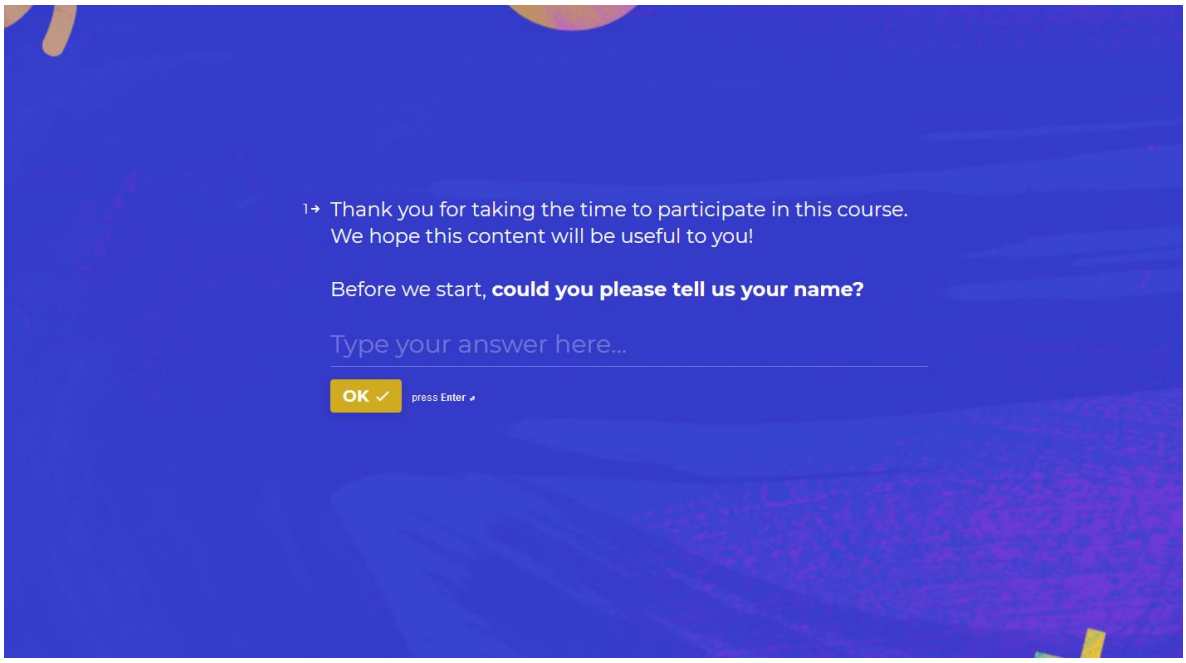

10→ Approximately, how much time did it take you to complete this course?
Type your answer here...
OK ✓
press Enter


11→ Are there any other materials pertaining to the annotation guidelines that you would have like to see in this course?
Type your answer here...
OK ✓
press Enter

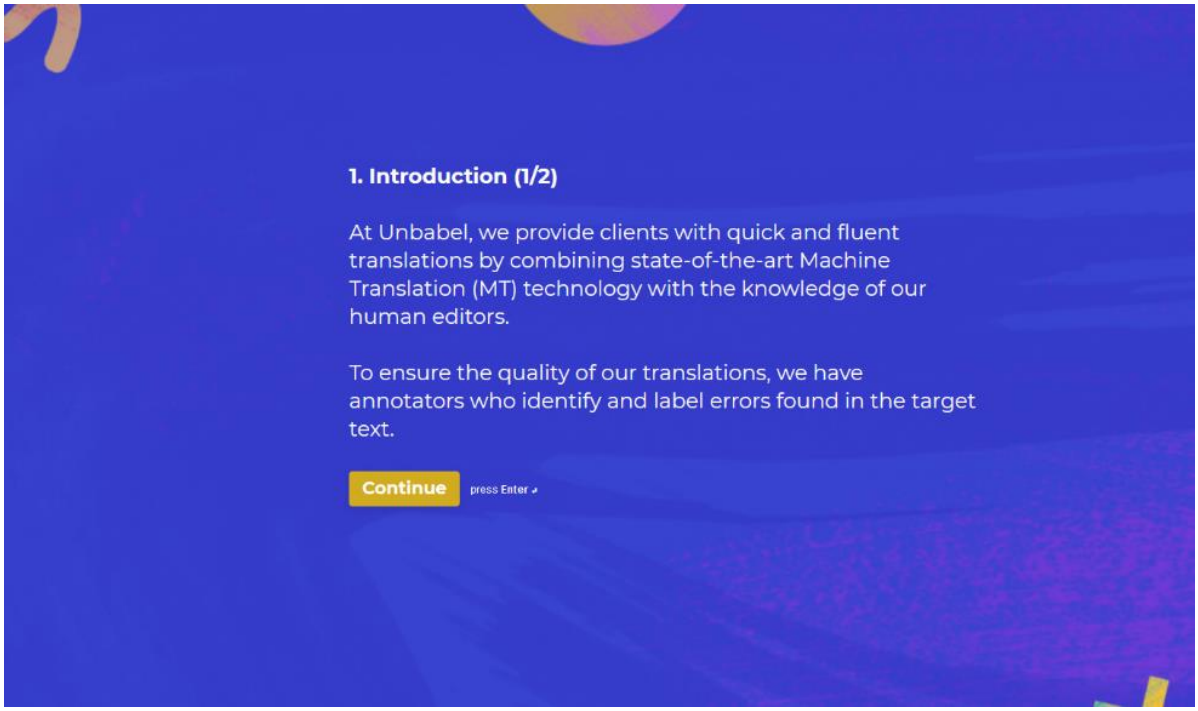

Thank you, [Name]!
These are your results:
You got 0/10 questions right.
See you next time!
Submit
press Ctrl + Enter

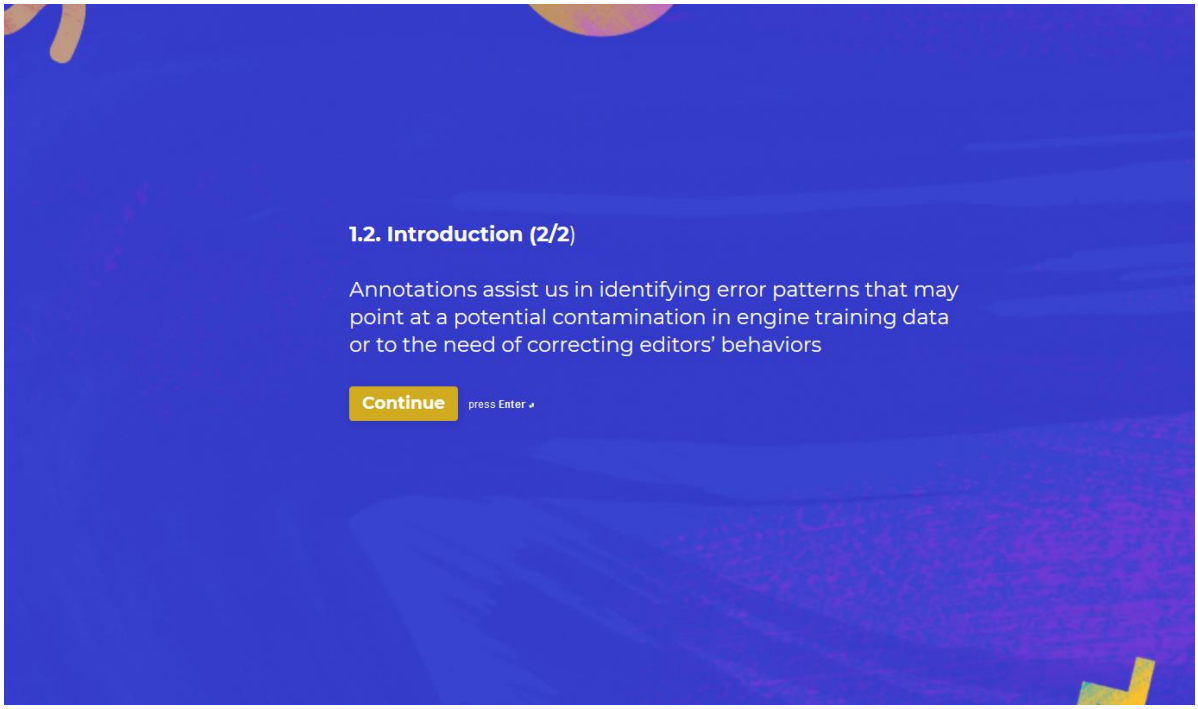
Annotation training (long version)[24]

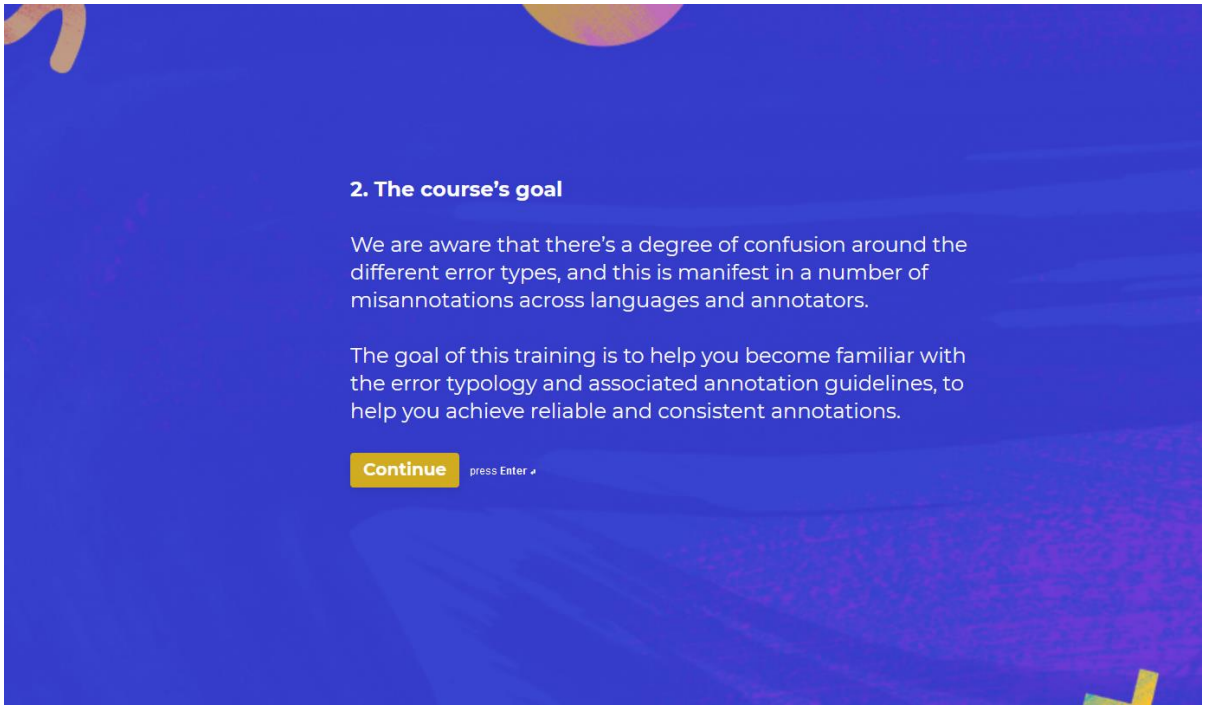


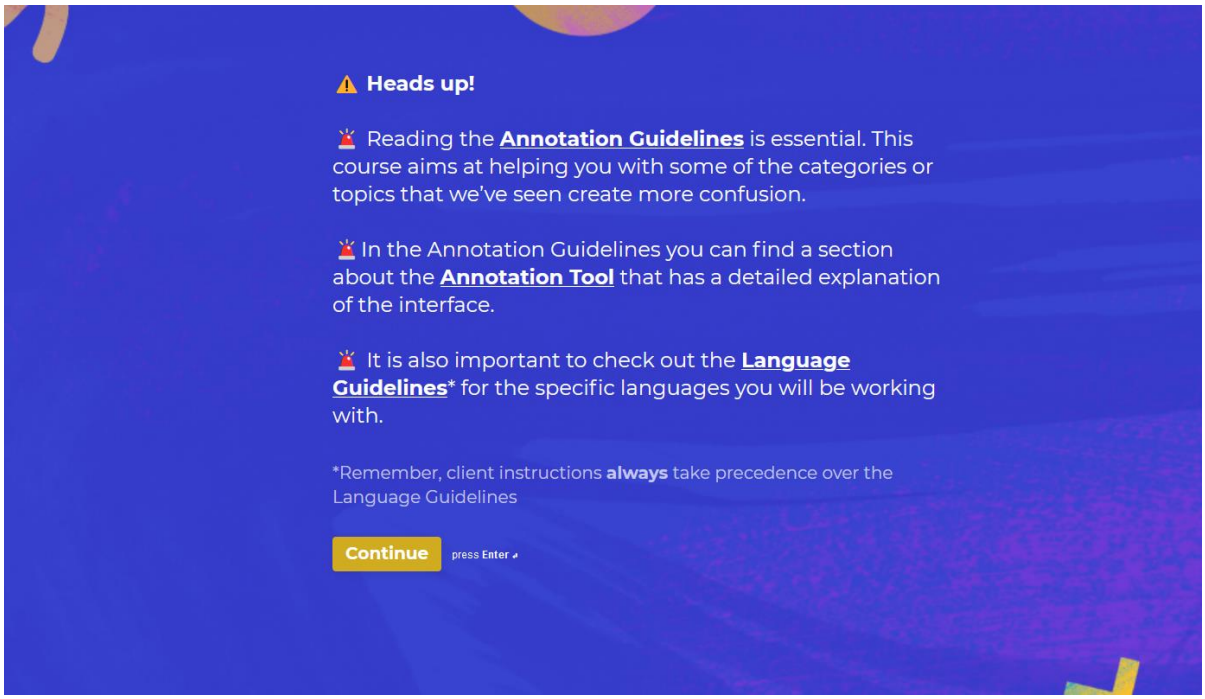


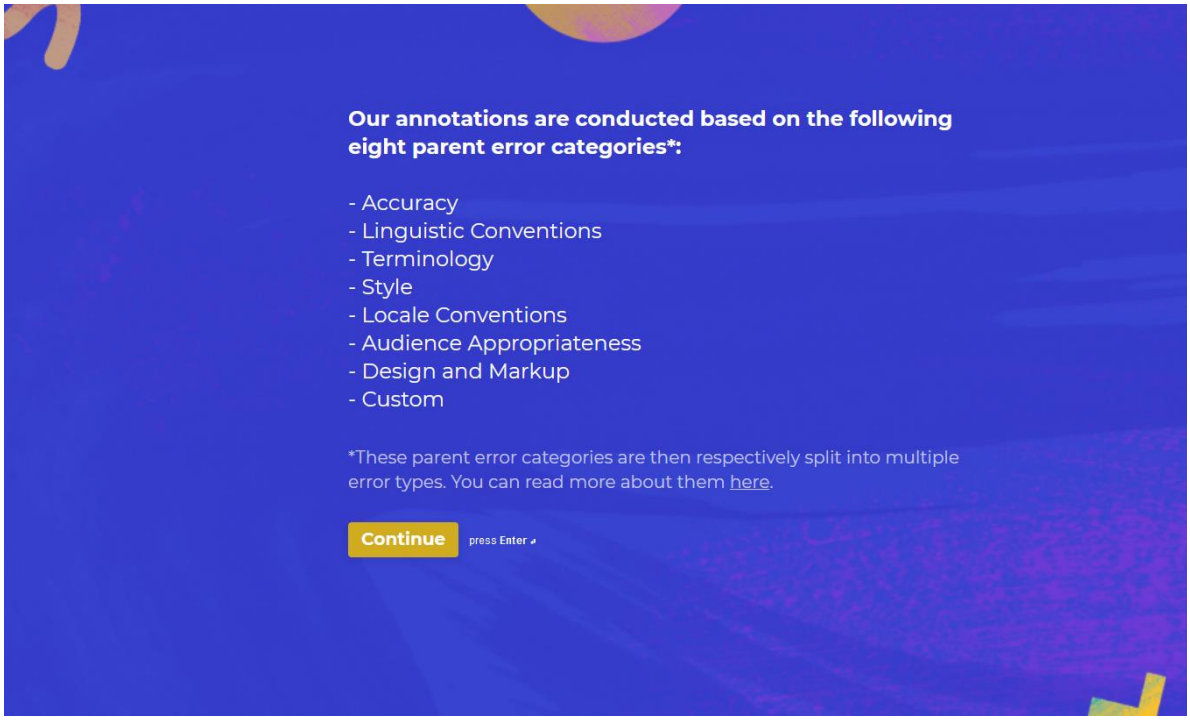
[24] The right answer for each question is selected. However, to demonstrate what annotators would see if they answered incorrectly, I answered every question incorrectly.

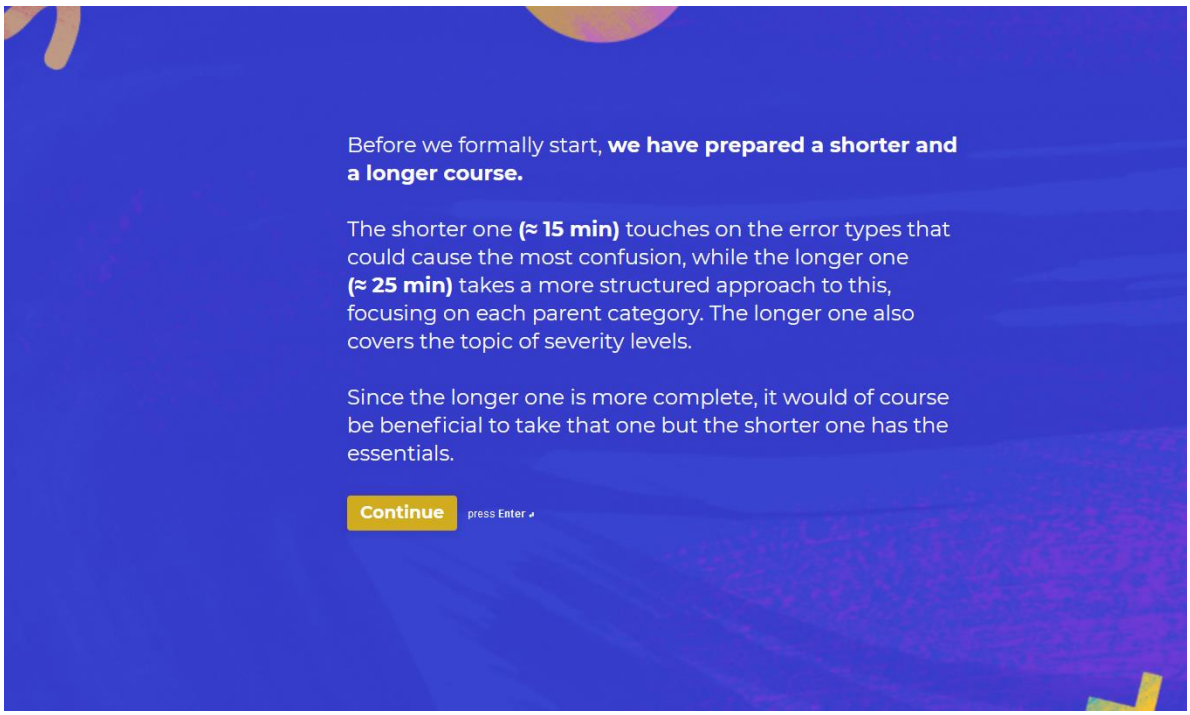

1. Introduction (1/2)
At Unbabel, we provide clients with quick and fluent translations by combining state-of-the-art Machine Translation (MT) technology with the knowledge of our human editors.
To ensure the quality of our translations, we have annotators who identify and label errors found in the target text.
Continue
press Enter

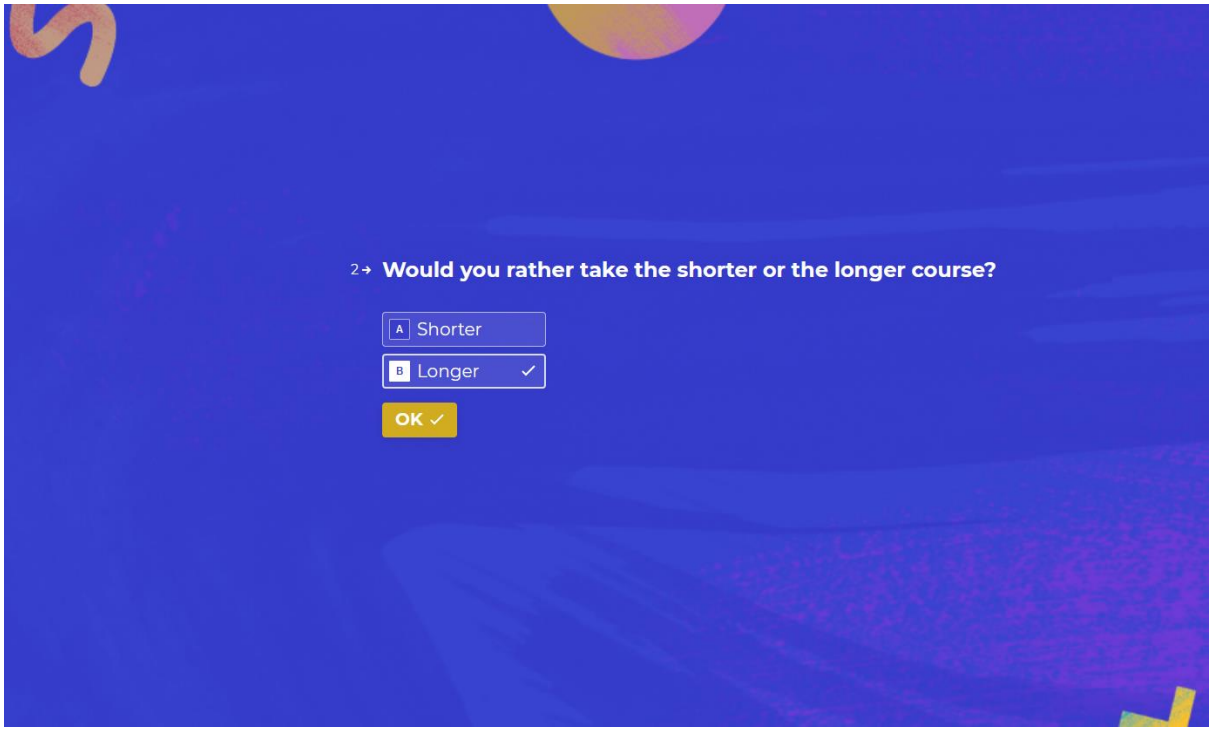

1.2. Introduction (2/2)
Annotations assist us in identifying error patterns that may point at a potential contamination in engine training data or to the need of correcting editors' behaviors
Continue
press Enter

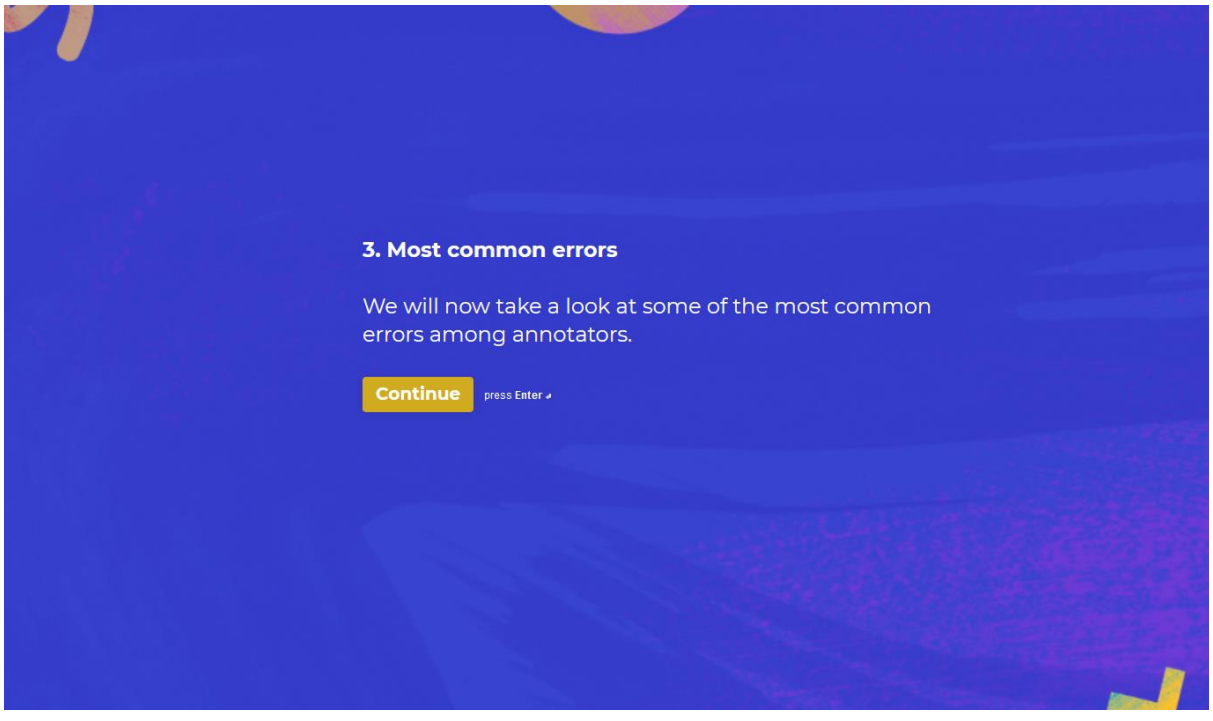

2. The course's goal
We are aware that there's a degree of confusion around the different error types, and this is manifest in a number of misannotations across languages and annotators.
The goal of this training is to help you become familiar with the error typology and associated annotation guidelines, to help you achieve reliable and consistent annotations.
Continue
press Enter

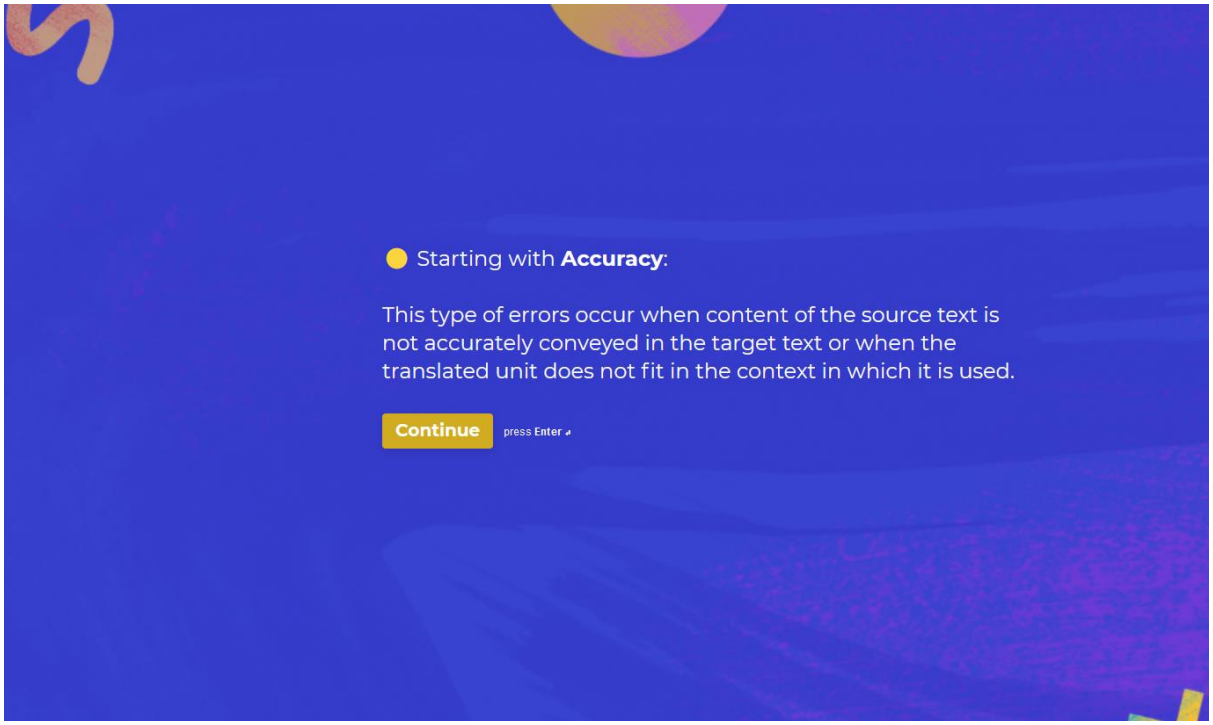

⚠ Heads up!
Reading the Annotation Guidelines is essential. This course aims at helping you with some of the categories or topics that we've seen create more confusion.
In the Annotation Guidelines you can find a section about the Annotation Tool that has a detailed explanation of the interface.
It is also important to check out the Language Guidelines* for the specific languages you will be working with.
*Remember, client instructions always take precedence over the Language Guidelines
Continue
press Enter

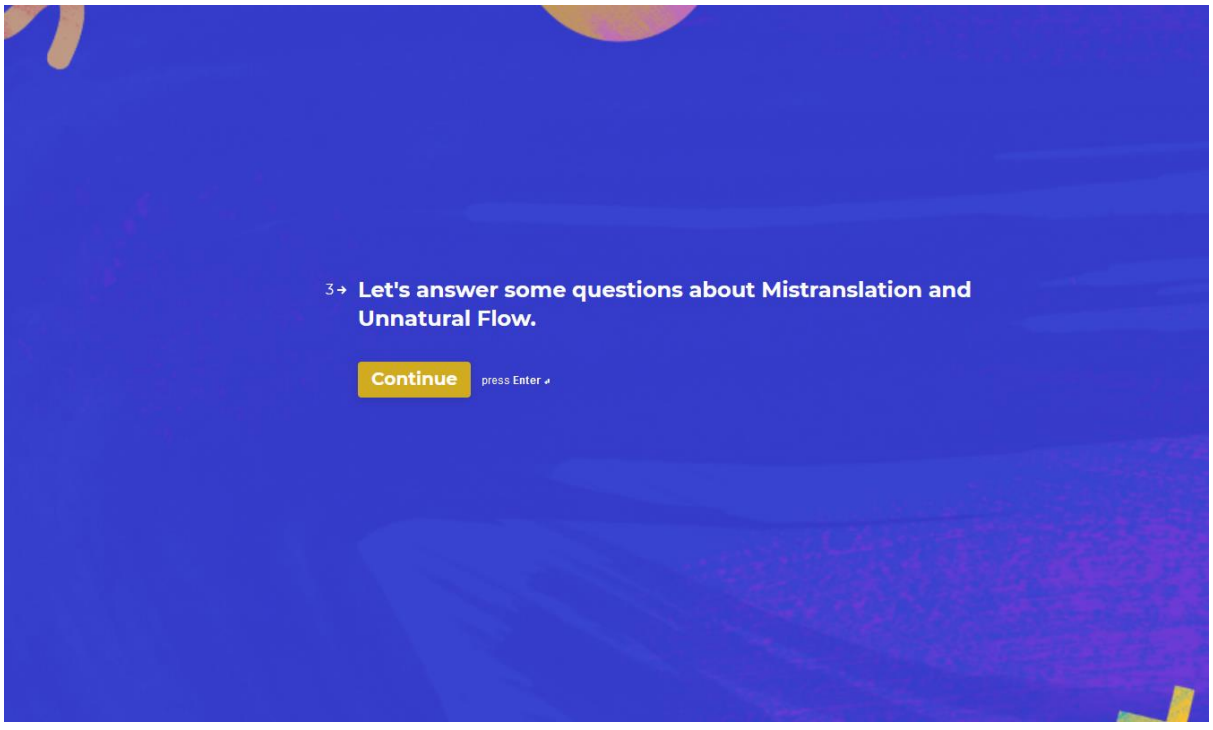

Our annotations are conducted based on the following eight parent error categories*:
- Accuracy
- Linguistic Conventions
- Terminology
- Style
- Locale Conventions
- Audience Appropriateness
- Design and Markup
- Custom
*These parent error categories are then respectively split into multiple error types. You can read more about them here.
Continue
press Enter


Before we formally start, we have prepared a shorter and a longer course.
The shorter one (≈ 15 min) touches on the error types that could cause the most confusion, while the longer one (≈ 25 min) takes a more structured approach to this, focusing on each parent category. The longer one also covers the topic of severity levels.
Since the longer one is more complete, it would of course be beneficial to take that one but the shorter one has the essentials.
Continue
press Enter


2→ Would you rather take the shorter or the longer course?
A Shorter
B Longer ✓
OK ✓


3. Most common errors
We will now take a look at some of the most common errors among annotators.
Continue
press Enter ↵


Starting with **Accuracy**:
This type of errors occur when content of the source text is not accurately conveyed in the target text or when the translated unit does not fit in the context in which it is used.
Continue
press Enter


3→ Let's answer some questions about Mistranslation and Unnatural Flow.
Continue
press Enter

a. How would you annotate this error?

Source [EN]	Target with annotation [PT-PT]	Context
"(...) thereby demonstrating its capacity to engineer a complex first of a kind large scale electrolysis-based ammonia project."	"(...) demonstrando assim a sua capacidade para conceber um complexo projeto de amoníaco baseado em eletrólise em larga escala, o primeiro do seu género."	While "o primeiro do seu género" is technically not incorrect, it is not something a native speaker would say as it sounds unnatural. In English, it would sound something like "(...) a complex first of its type large scale electrolysis-based ammonia project."

A Mistranslation

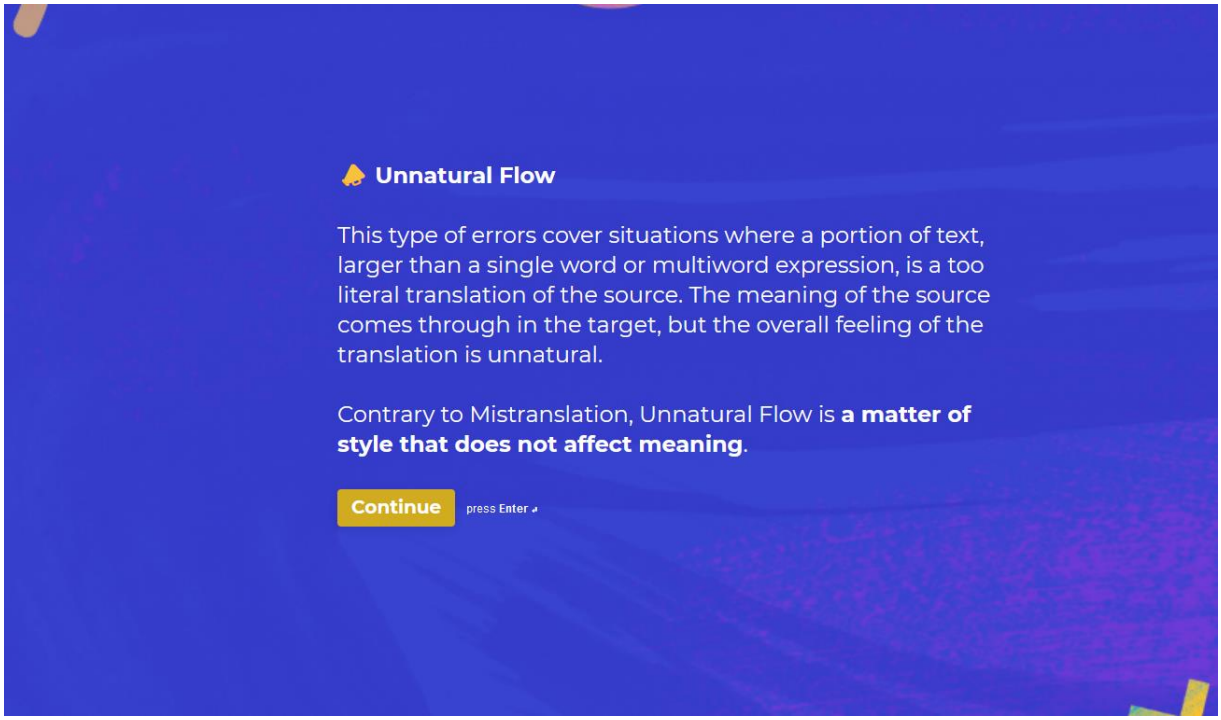
B Unnatural Flow ✓

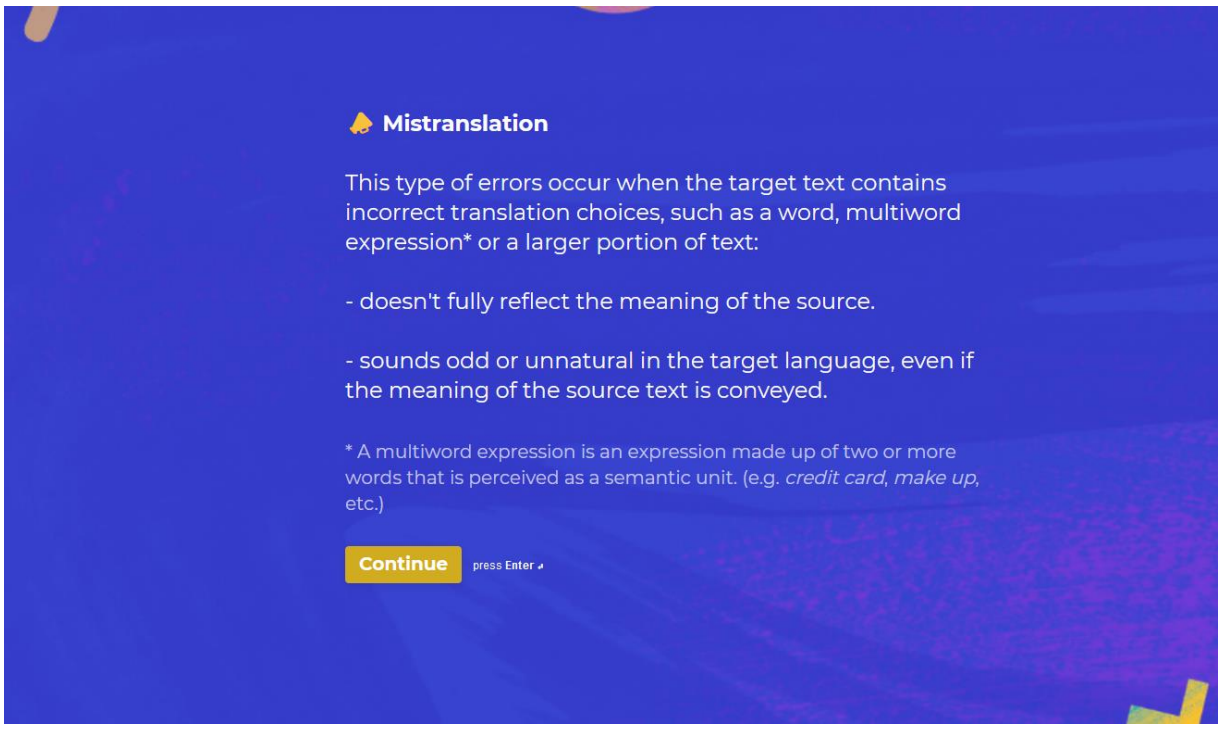
OK ✓

✗ **Hmmm... No.**

This should be marked as **Unnatural Flow**. Even though the meaning of the source text is conveyed in the translation, "o primeiro do seu género" sounds unnatural to a native reader.

Continue press Enter


Unnatural Flow
This type of errors cover situations where a portion of text, larger than a single word or multiword expression, is a too literal translation of the source. The meaning of the source comes through in the target, but the overall feeling of the translation is unnatural.
Contrary to Mistranslation, Unnatural Flow is a matter of style that does not affect meaning.
Continue
press Enter


Mistranslation
This type of errors occur when the target text contains incorrect translation choices, such as a word, multiword expression* or a larger portion of text:
- doesn't fully reflect the meaning of the source.
- sounds odd or unnatural in the target language, even if the meaning of the source text is conveyed.
* A multiword expression is an expression made up of two or more words that is perceived as a semantic unit. (e.g. credit card, make up, etc.)
Continue
press Enter

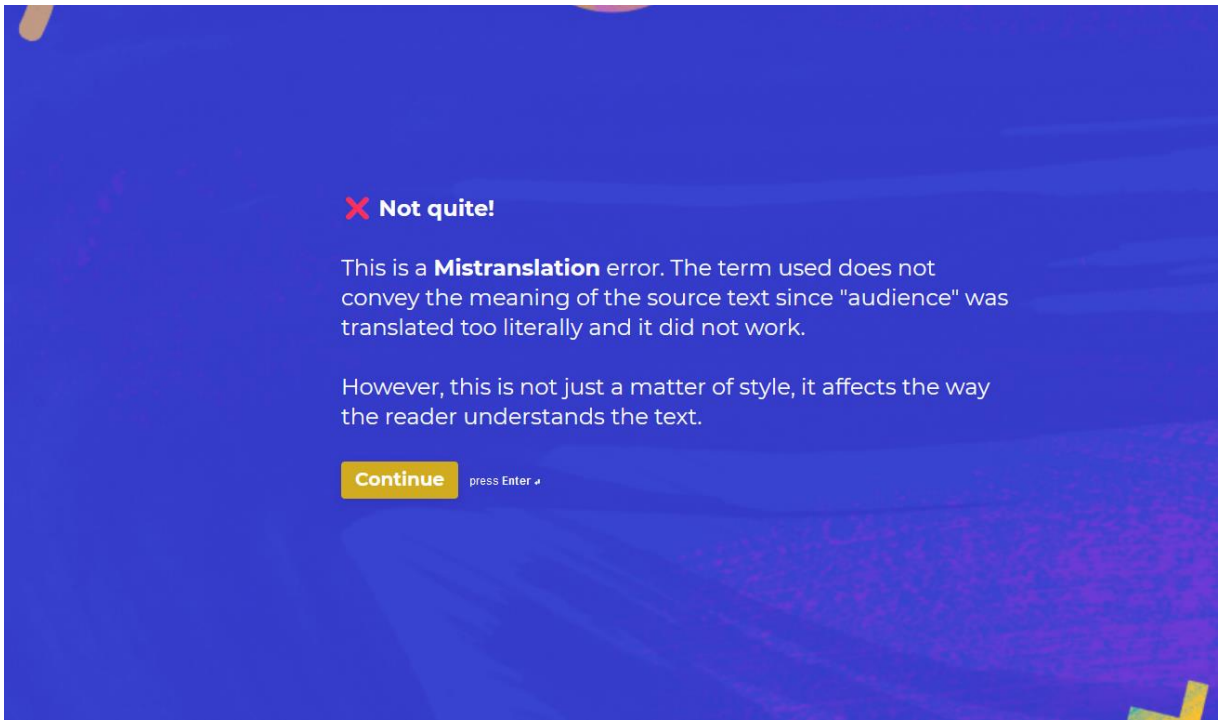
b. How would you annotate this error?

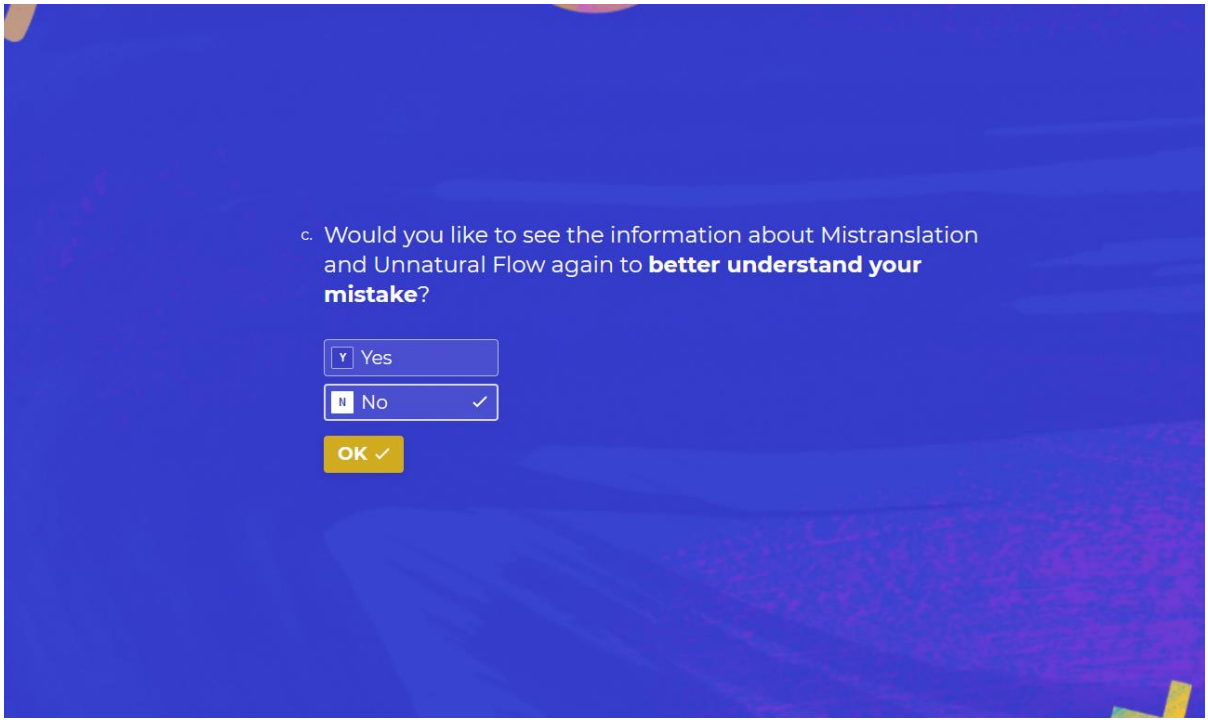
Source [EN]	Target with annotation [PT-PT]	Context
"Our goal is to bring elements that are synonymous with the Netherlands to an International audience."	"O nosso objetivo é possibilitar a uma audiência internacional desfrutar de elementos que são sinónimo dos Países Baixos."	"audience" is translated too literally, it is too close to the target text. In Portuguese, one would use audiência more as "hearing 99 (i.e. court hearing).

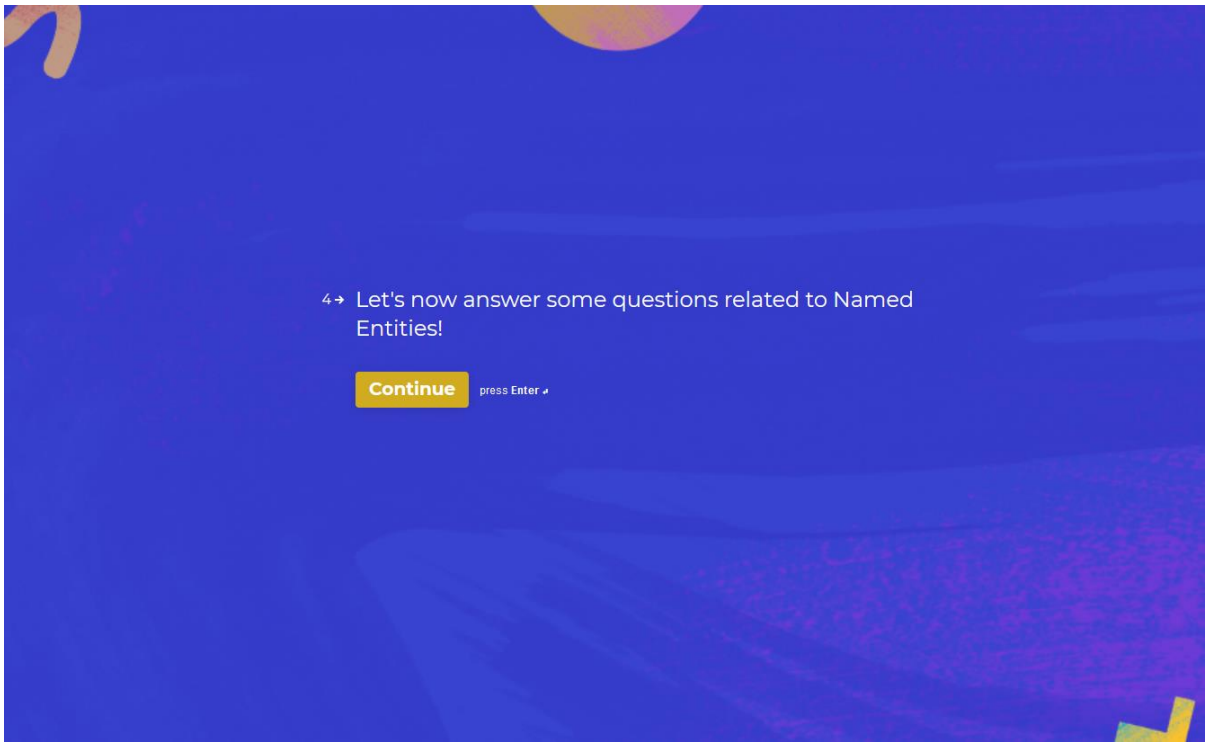
A Mistranslation ✓

B Unnatural Flow

OK ✓




c. Would you like to see the information about Mistranslation and Unnatural Flow again to **better understand your mistake**?
Y Yes
N No ✓
OK ✓


4→ Let's now answer some questions related to Named Entities!
Continue
press Enter

a. How would you annotate this error?

Source [EN]	Target with annotation [PT-PT]	Context
"It has a starting price of $49.99 a year."	"Tem um preço inicial de $49.99 por ano."	In Portuguese you'd translate it as "49,99 $" (dollar sign in the wrong place, period instead of comma and missing whitespace).

A Untranslated

B Wrong Named Entity

C Currency Format ✓

OK ✓

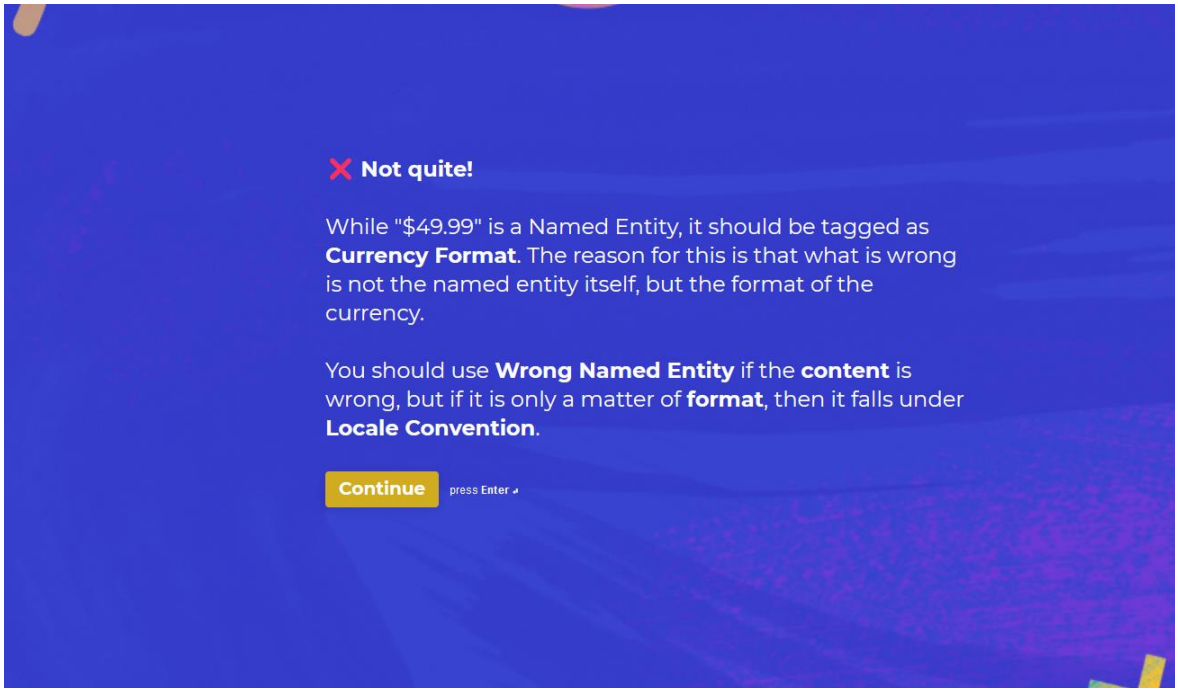
❌ **Not really...**

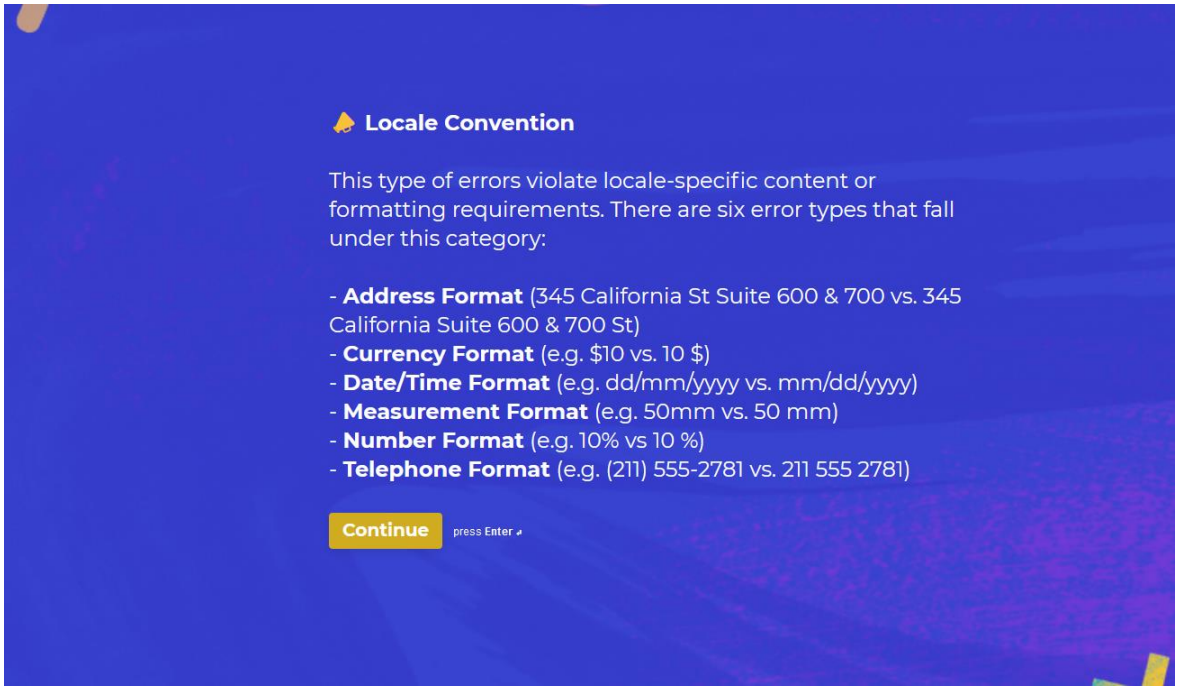
When named entities are left untranslated (i.e. a date), it may be tempting to tag this as **Untranslated**, but it would still be considered **Wrong Named Entity**.

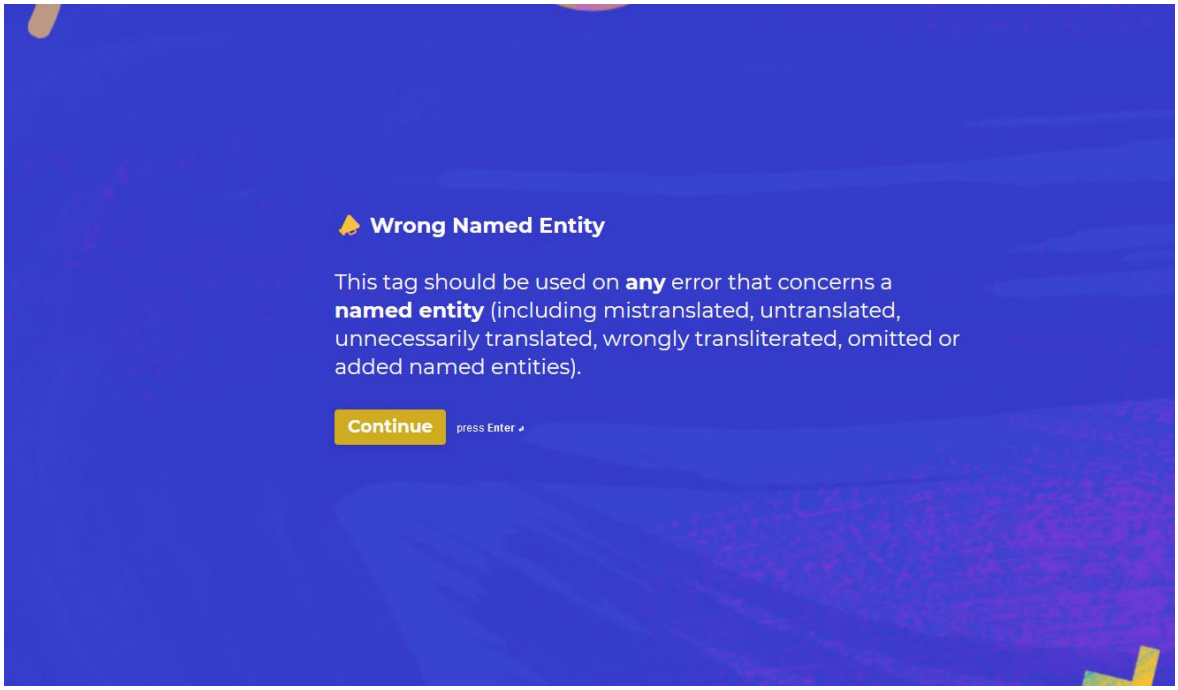
Even so, while "$49.99" is a named entity, it should be tagged as **Currency Format**. The reason for this is that what is wrong is not the named entity itself, but the format of the currency.

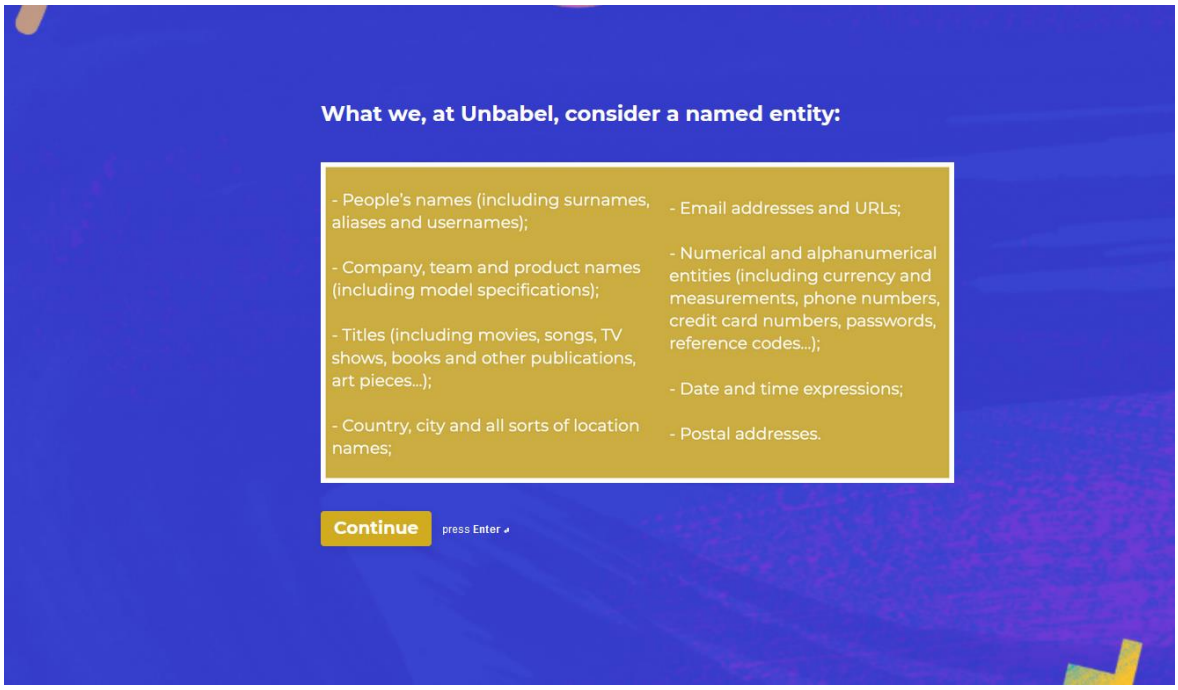
You should use **Wrong Named Entity** if the content is wrong, but if it is only a matter of format, then it falls under **Locale Convention**.

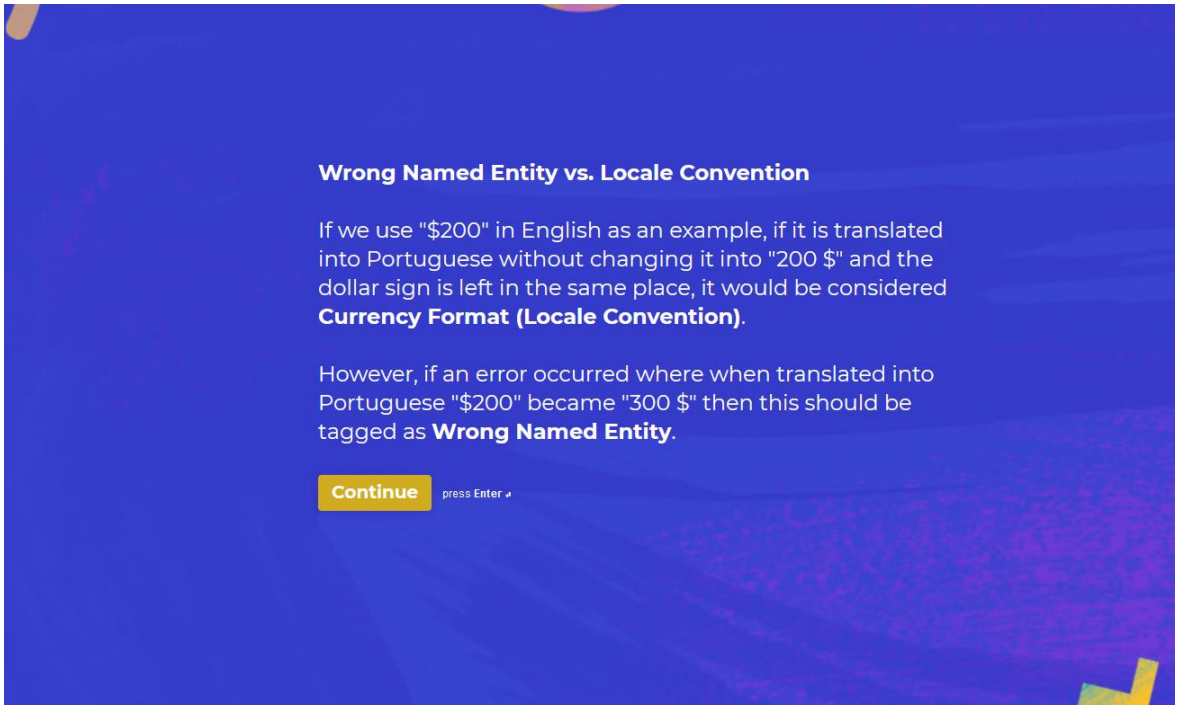
Continue press Enter

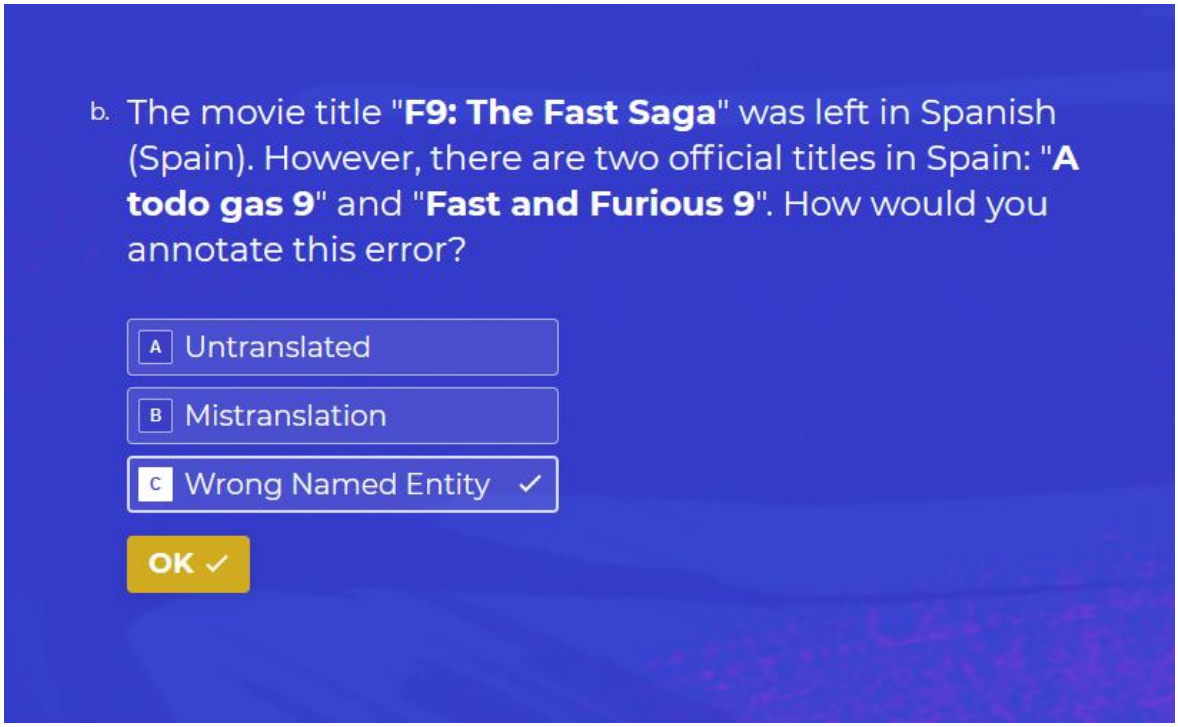

❌ Not quite!
While "$49.99" is a Named Entity, it should be tagged as **Currency Format**. The reason for this is that what is wrong is not the named entity itself, but the format of the currency.
You should use **Wrong Named Entity** if the **content** is wrong, but if it is only a matter of **format**, then it falls under **Locale Convention**.
Continue
press Enter

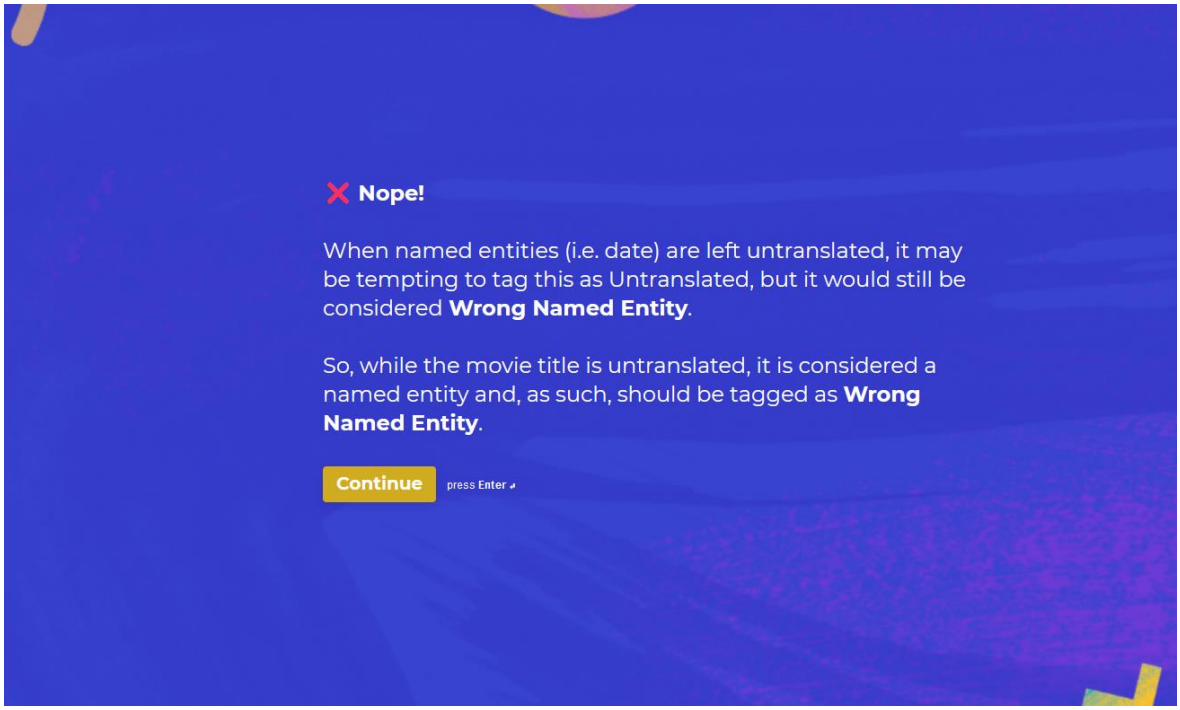

🔔 Locale Convention
This type of errors violate locale-specific content or formatting requirements. There are six error types that fall under this category:
- **Address Format** (345 California St Suite 600 & 700 vs. 345 California Suite 600 & 700 St)
- **Currency Format** (e.g. $10 vs. 10 $)
- **Date/Time Format** (e.g. dd/mm/yyyy vs. mm/dd/yyyy)
- **Measurement Format** (e.g. 50mm vs. 50 mm)
- **Number Format** (e.g. 10% vs 10 %)
- **Telephone Format** (e.g. (211) 555-2781 vs. 211 555 2781)
Continue
press Enter

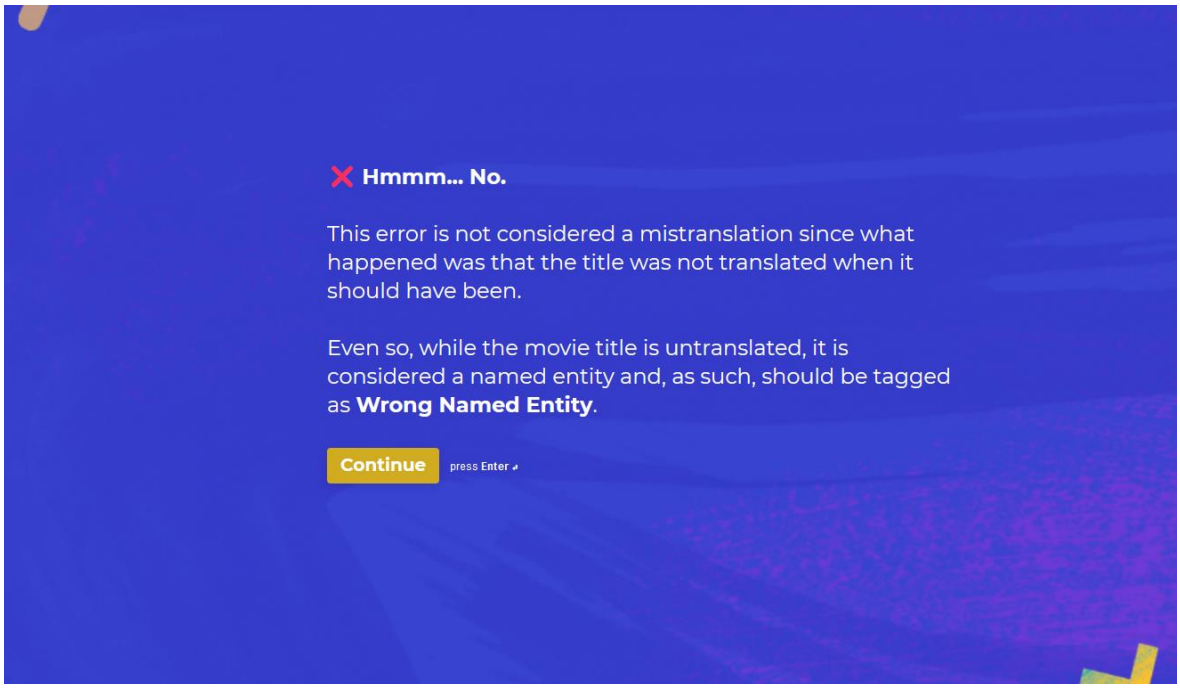

Wrong Named Entity
This tag should be used on **any** error that concerns a **named entity** (including mistranslated, untranslated, unnecessarily translated, wrongly transliterated, omitted or added named entities).
Continue
press Enter

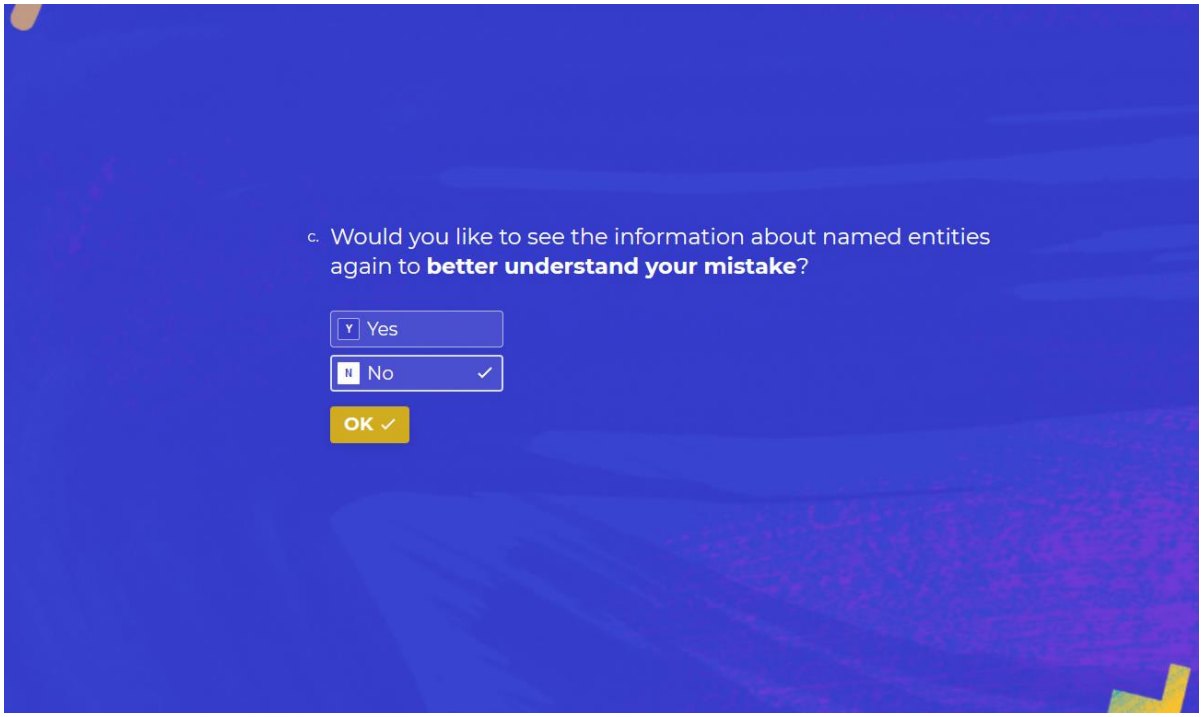

What we, at Unbabel, consider a named entity:
- People's names (including surnames, aliases and usernames);
- Company, team and product names (including model specifications);
- Titles (including movies, songs, TV shows, books and other publications, art pieces...);
- Country, city and all sorts of location names;
- Email addresses and URLs;
- Numerical and alphanumerical entities (including currency and measurements, phone numbers, credit card numbers, passwords, reference codes...);
- Date and time expressions;
- Postal addresses.
Continue
press Enter

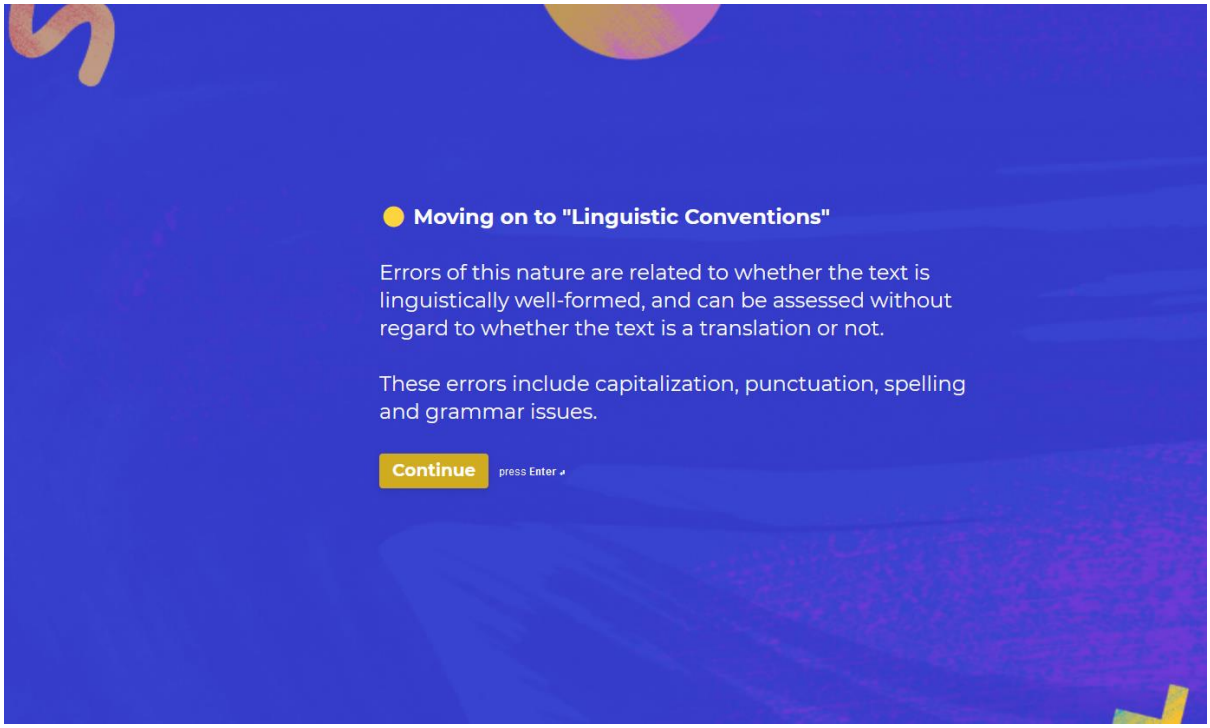

Wrong Named Entity vs. Locale Convention
If we use "$200" in English as an example, if it is translated into Portuguese without changing it into "200 $" and the dollar sign is left in the same place, it would be considered Currency Format (Locale Convention).
However, if an error occurred where when translated into Portuguese "$200" became "300 $" then this should be tagged as Wrong Named Entity.
Continue
press Enter

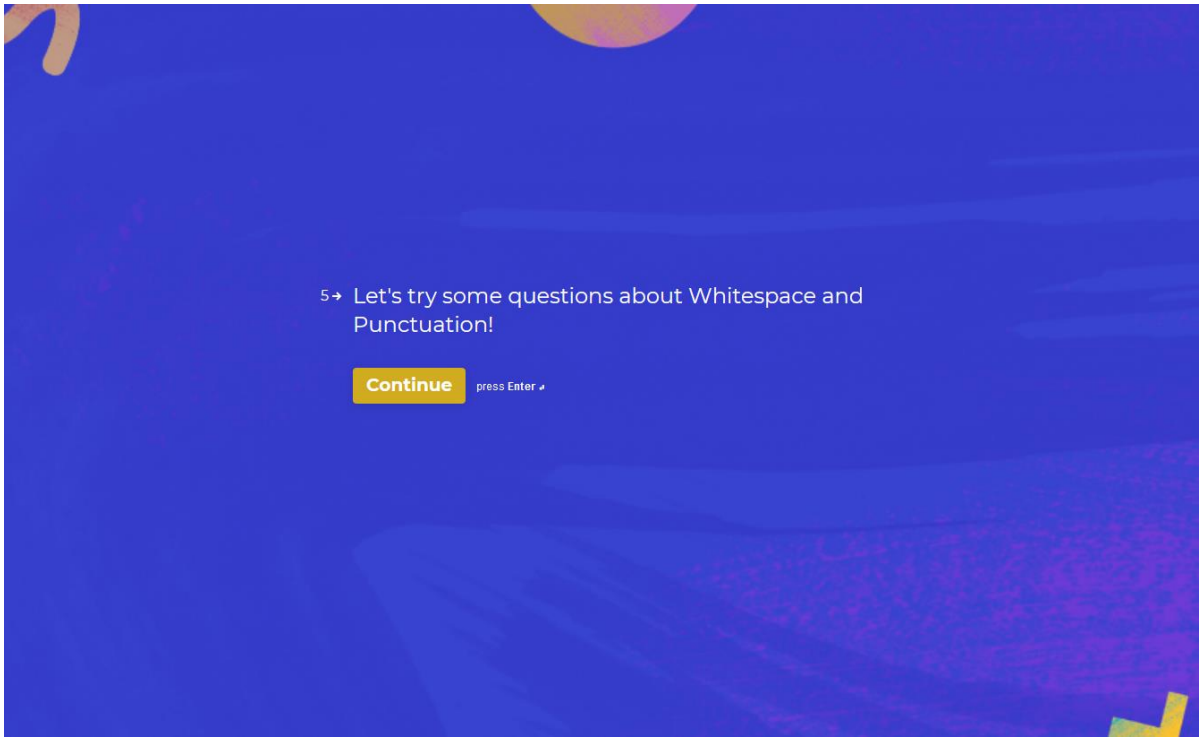

b. The movie title "F9: The Fast Saga" was left in Spanish (Spain). However, there are two official titles in Spain: "A todo gas 9" and "Fast and Furious 9". How would you annotate this error?
A Untranslated
B Mistranslation
C Wrong Named Entity
OK


Nope!
When named entities (i.e. date) are left untranslated, it may be tempting to tag this as Untranslated, but it would still be considered Wrong Named Entity.
So, while the movie title is untranslated, it is considered a named entity and, as such, should be tagged as Wrong Named Entity.
Continue
press Enter


Hmmm... No.
This error is not considered a mistranslation since what happened was that the title was not translated when it should have been.
Even so, while the movie title is untranslated, it is considered a named entity and, as such, should be tagged as Wrong Named Entity.
Continue
press Enter


c. Would you like to see the information about named entities again to **better understand your mistake**?
Y Yes
N No
OK


Moving on to "Linguistic Conventions"
Errors of this nature are related to whether the text is linguistically well-formed, and can be assessed without regard to whether the text is a translation or not.
These errors include capitalization, punctuation, spelling and grammar issues.
Continue
press Enter



a. How would you annotate this error?

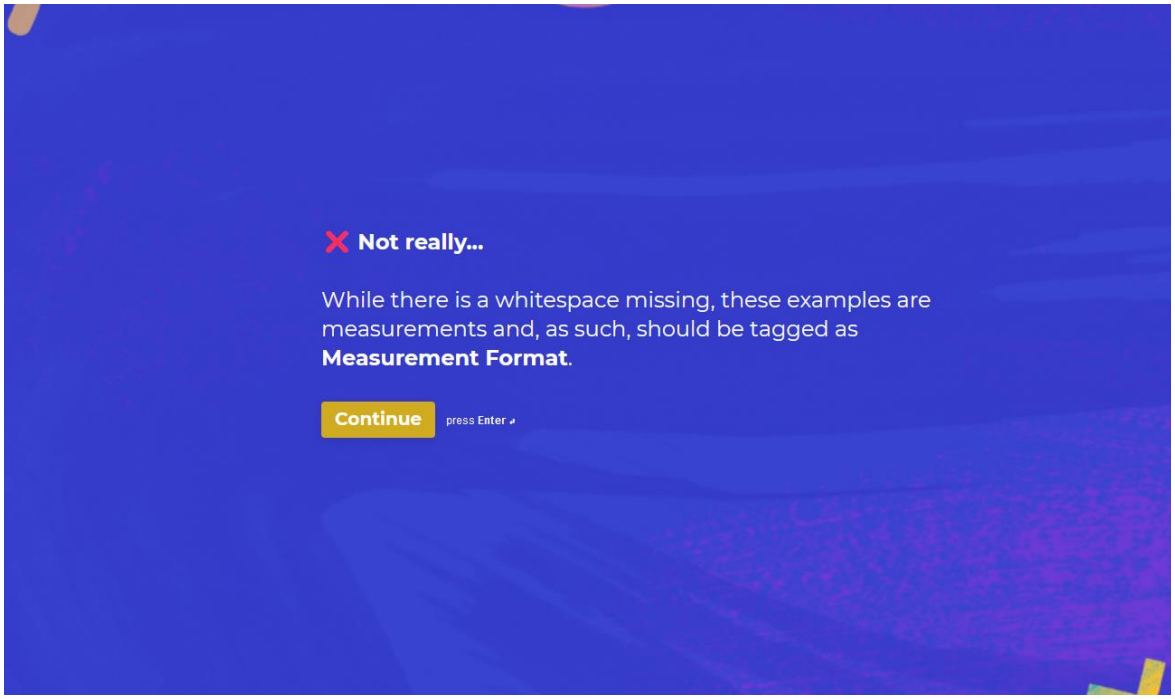
Source [EN]	Target with annotation [SV]	Context
"A 4.8V threshold voltage that allows 0V to 15V drive."	"På 4.8V som möjliggör drivenhet mellan 0V och 15V."	In Swedish, there should be a space between the number and the 'V'.

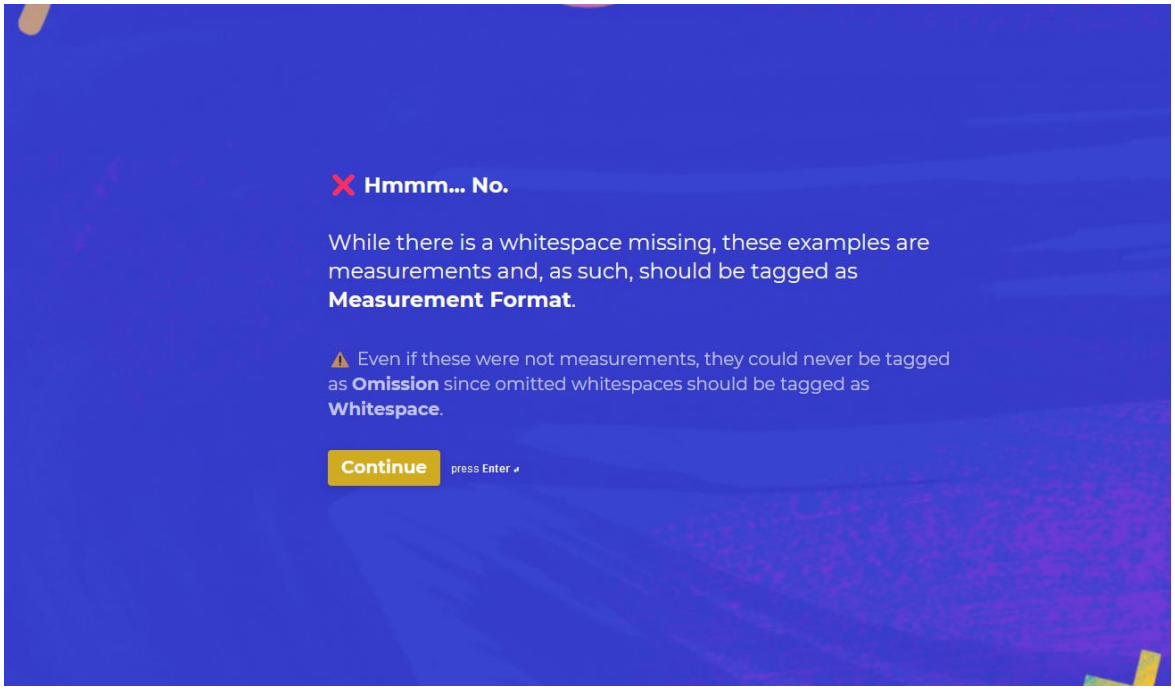
A Whitespace

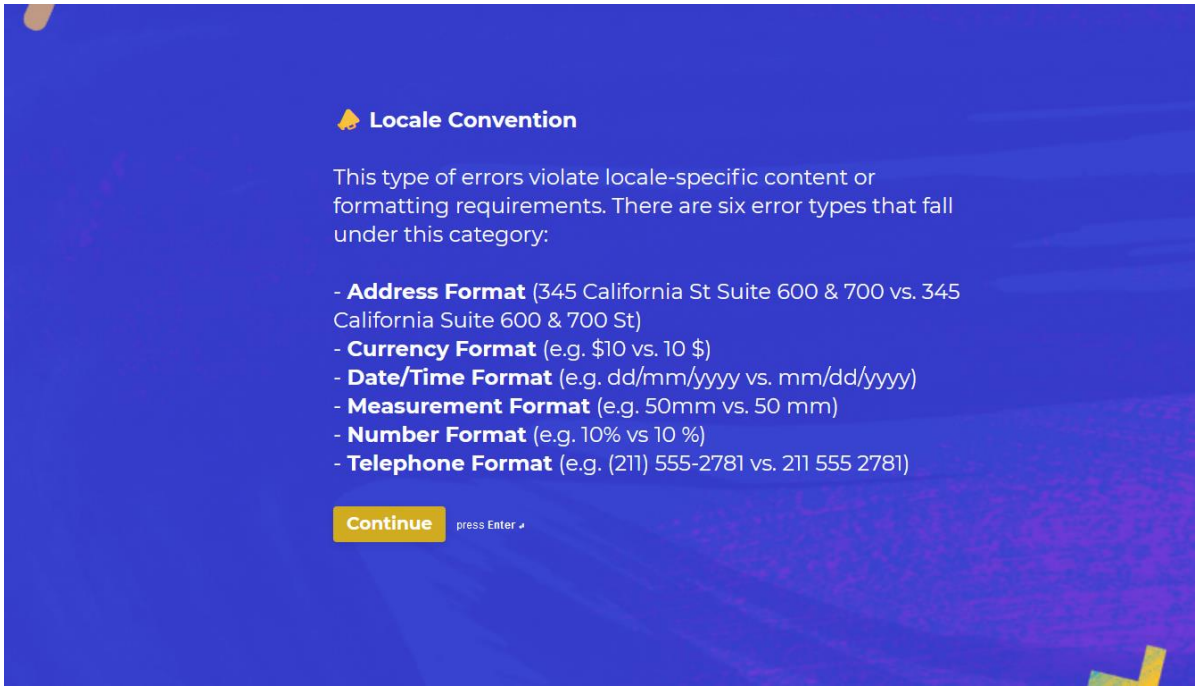
B Omission

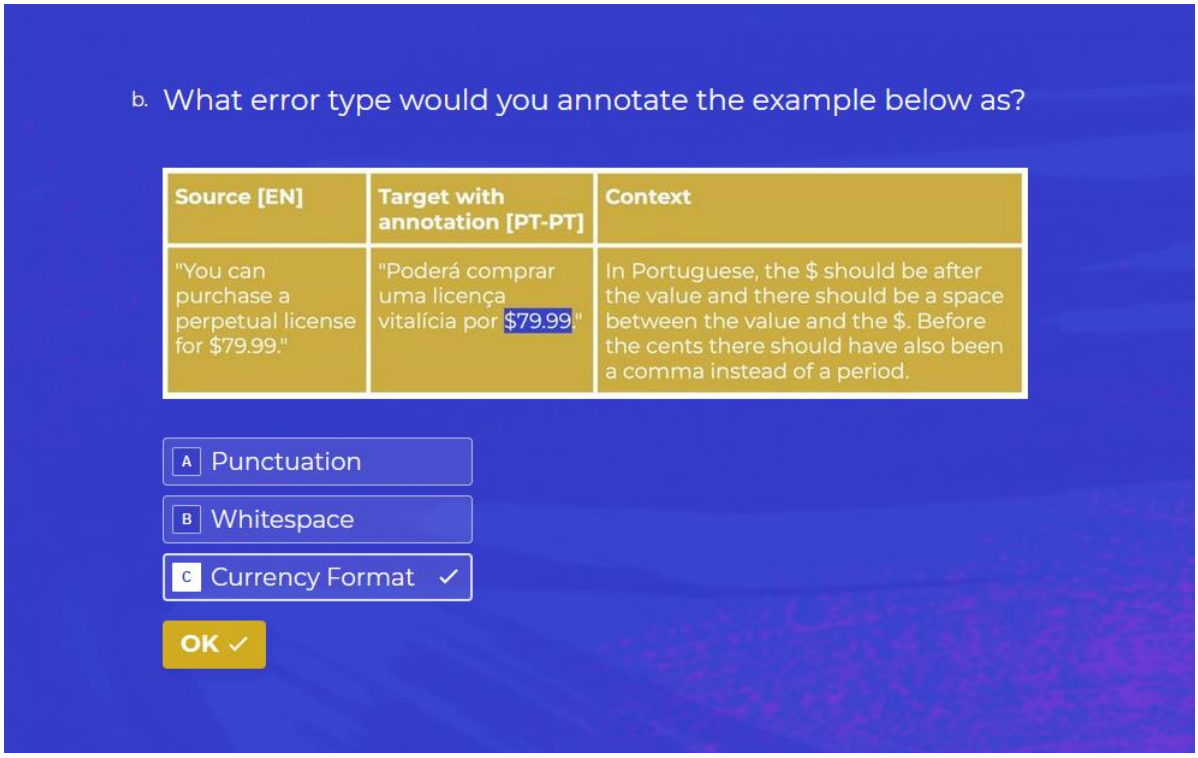
C Measurement Format ✓

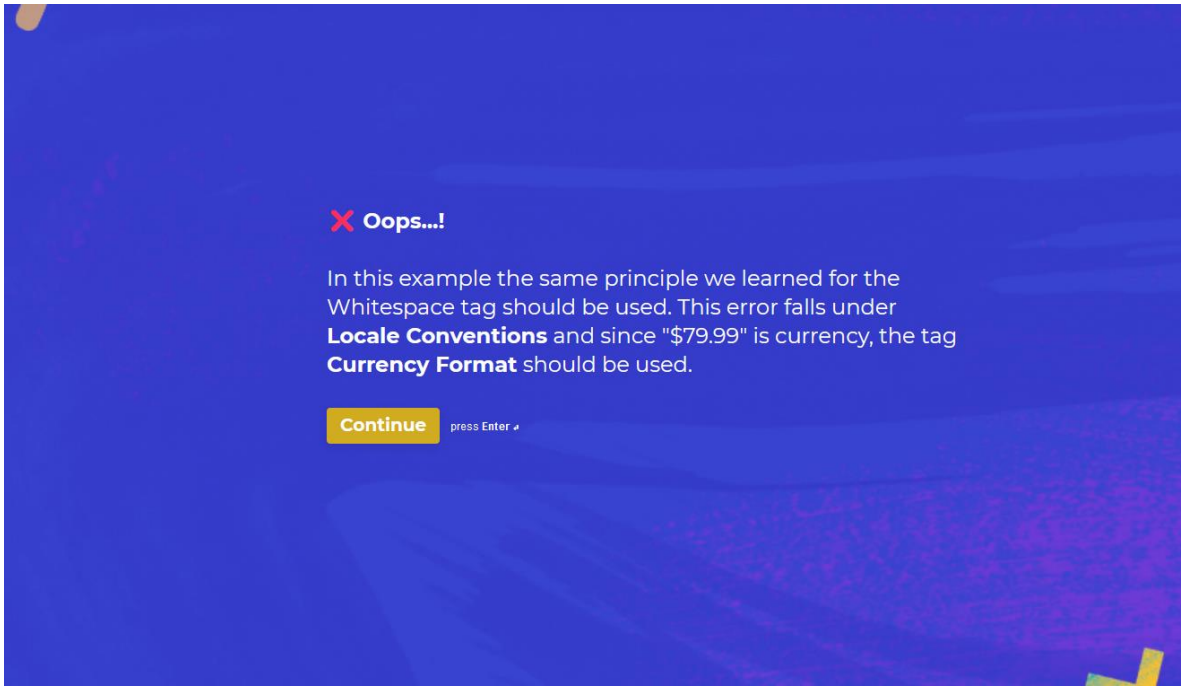
OK ✓

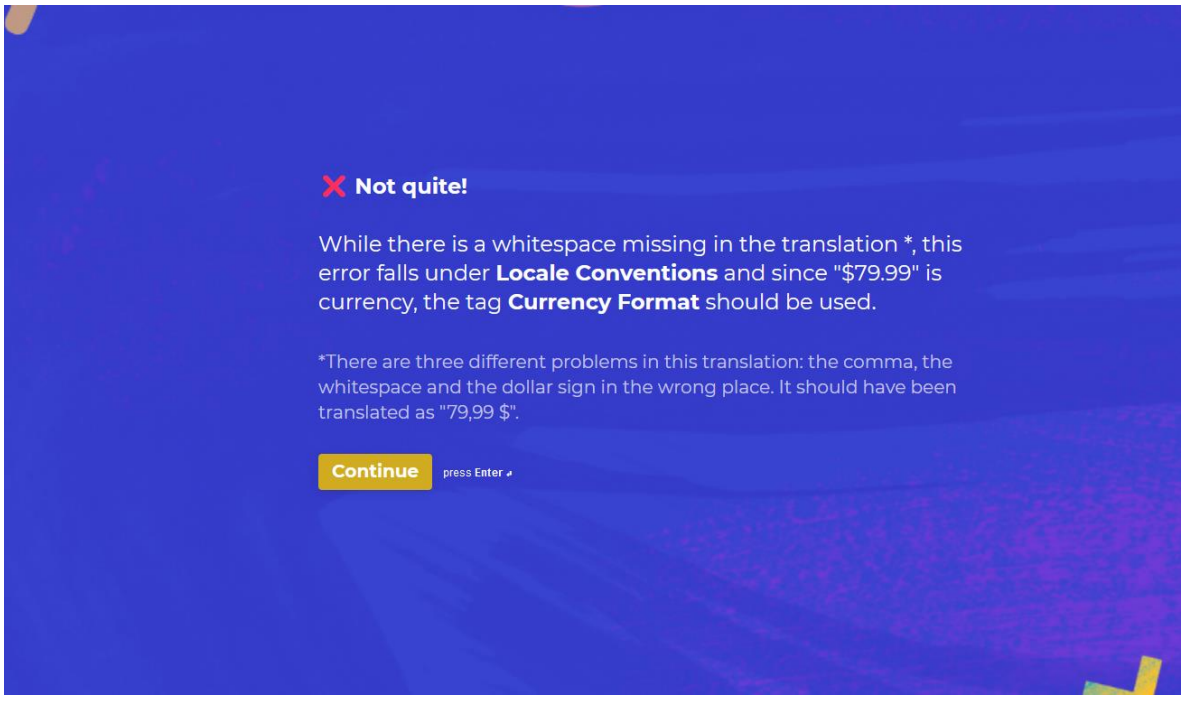

❌ Not really...
While there is a whitespace missing, these examples are measurements and, as such, should be tagged as **Measurement Format**.
Continue
press Enter

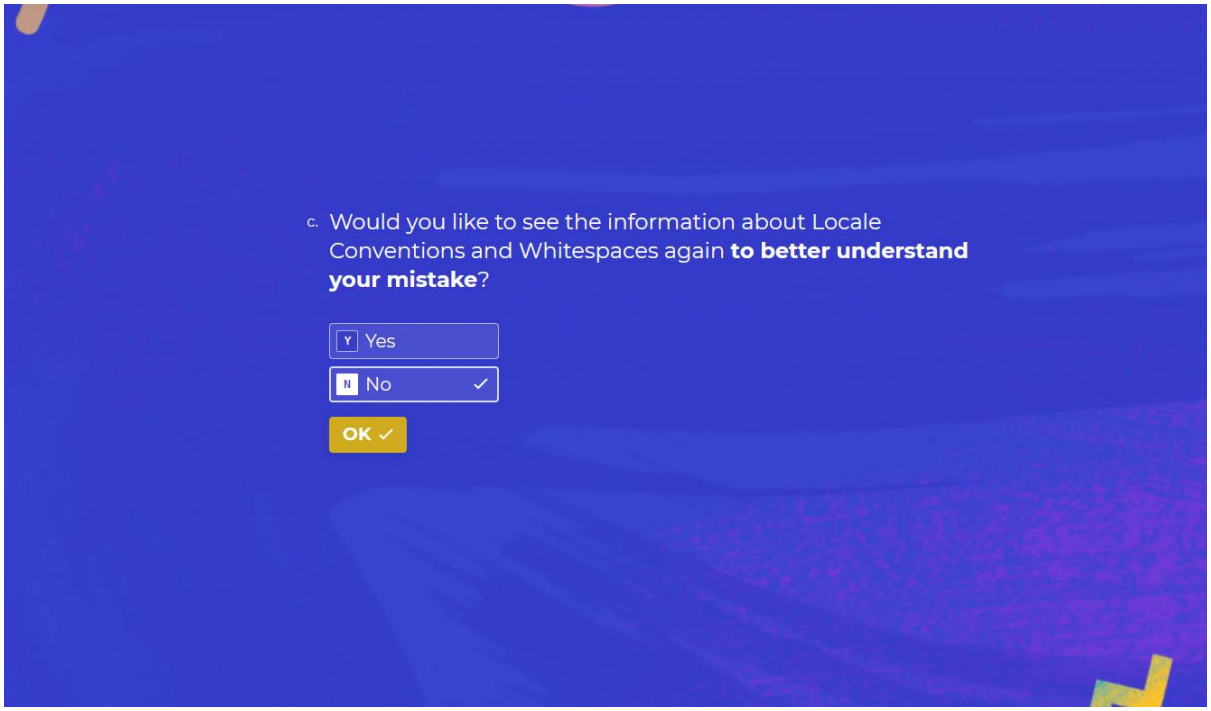

❌ Hmmm... No.
While there is a whitespace missing, these examples are measurements and, as such, should be tagged as **Measurement Format**.
⚠ Even if these were not measurements, they could never be tagged as **Omission** since omitted whitespaces should be tagged as **Whitespace**.
Continue
press Enter

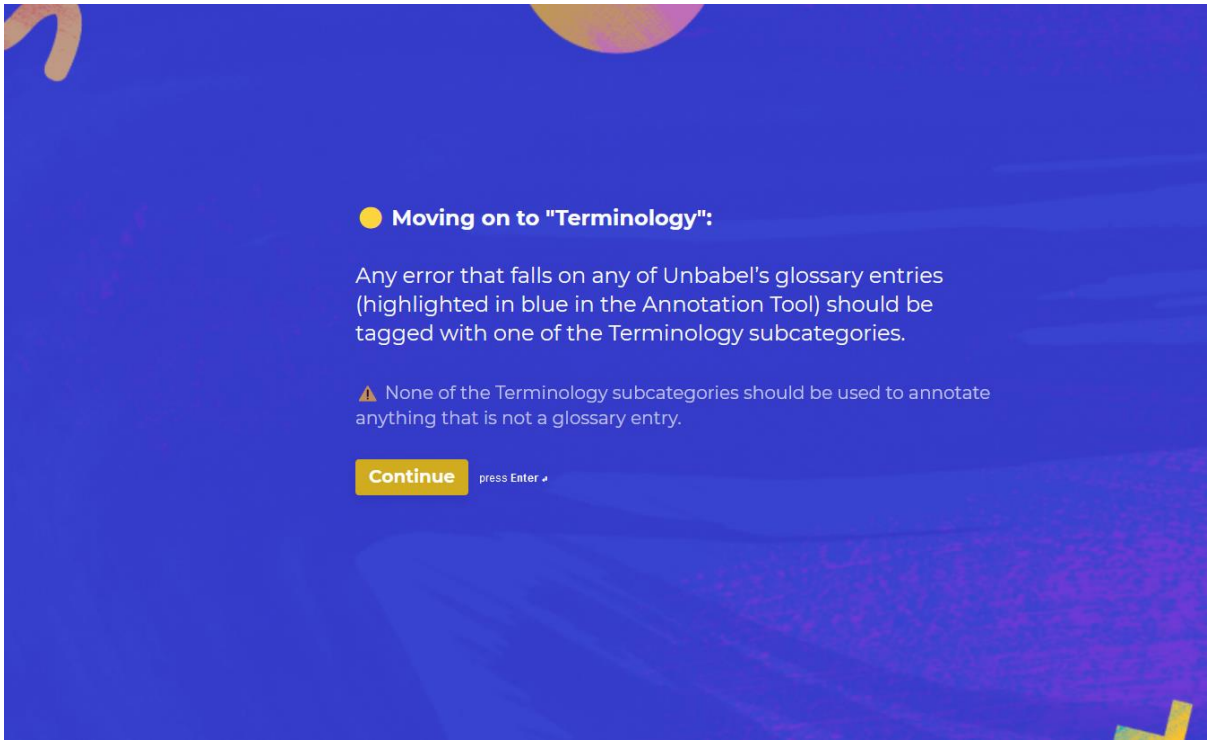

Locale Convention
This type of errors violate locale-specific content or formatting requirements. There are six error types that fall under this category:
- Address Format (345 California St Suite 600 & 700 vs. 345 California Suite 600 & 700 St)
- Currency Format (e.g. $10 vs. 10 $)
- Date/Time Format (e.g. dd/mm/yyyy vs. mm/dd/yyyy)
- Measurement Format (e.g. 50mm vs. 50 mm)
- Number Format (e.g. 10% vs 10 %)
- Telephone Format (e.g. (211) 555-2781 vs. 211 555 2781)
Continue
press Enter

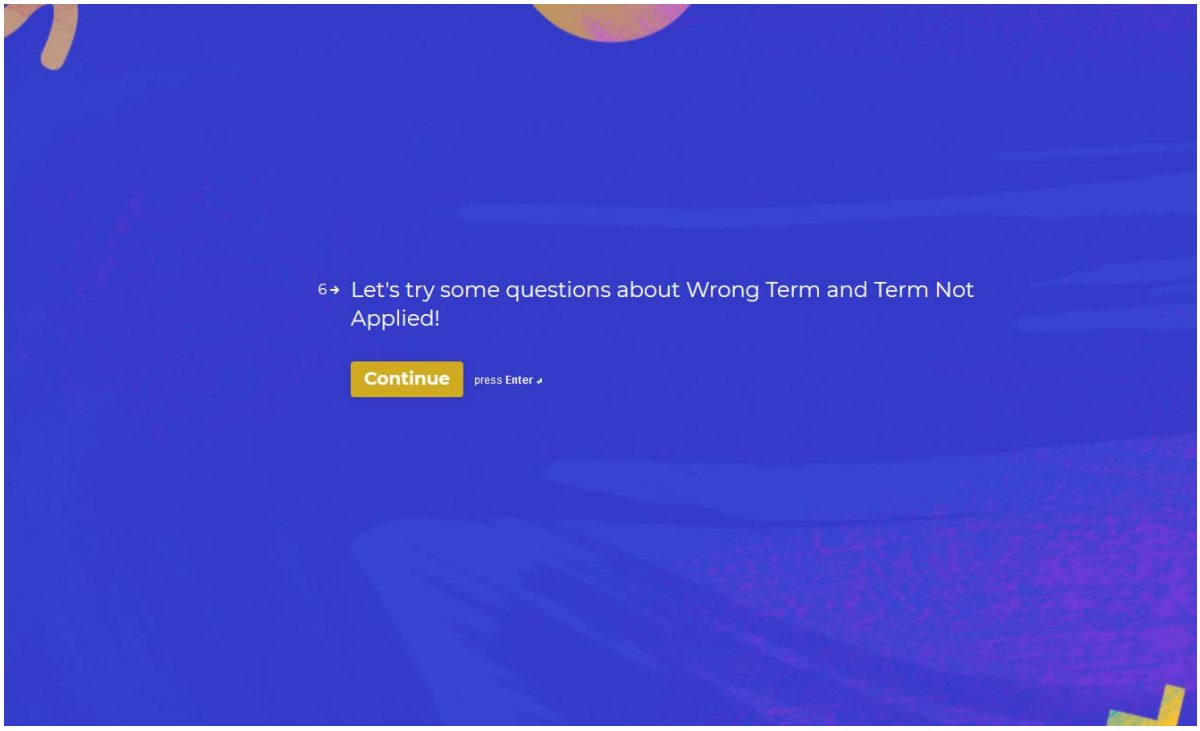

b. What error type would you annotate the example below as?
Source [EN]
Target with annotation [PT-PT]
Context
"You can purchase a perpetual license for $79.99."
"Poderá comprar uma licença vitalícia por $79.99"
In Portuguese, the $ should be after the value and there should be a space between the value and the $. Before the cents there should have also been a comma instead of a period.
A Punctuation
B Whitespace
C Currency Format
OK

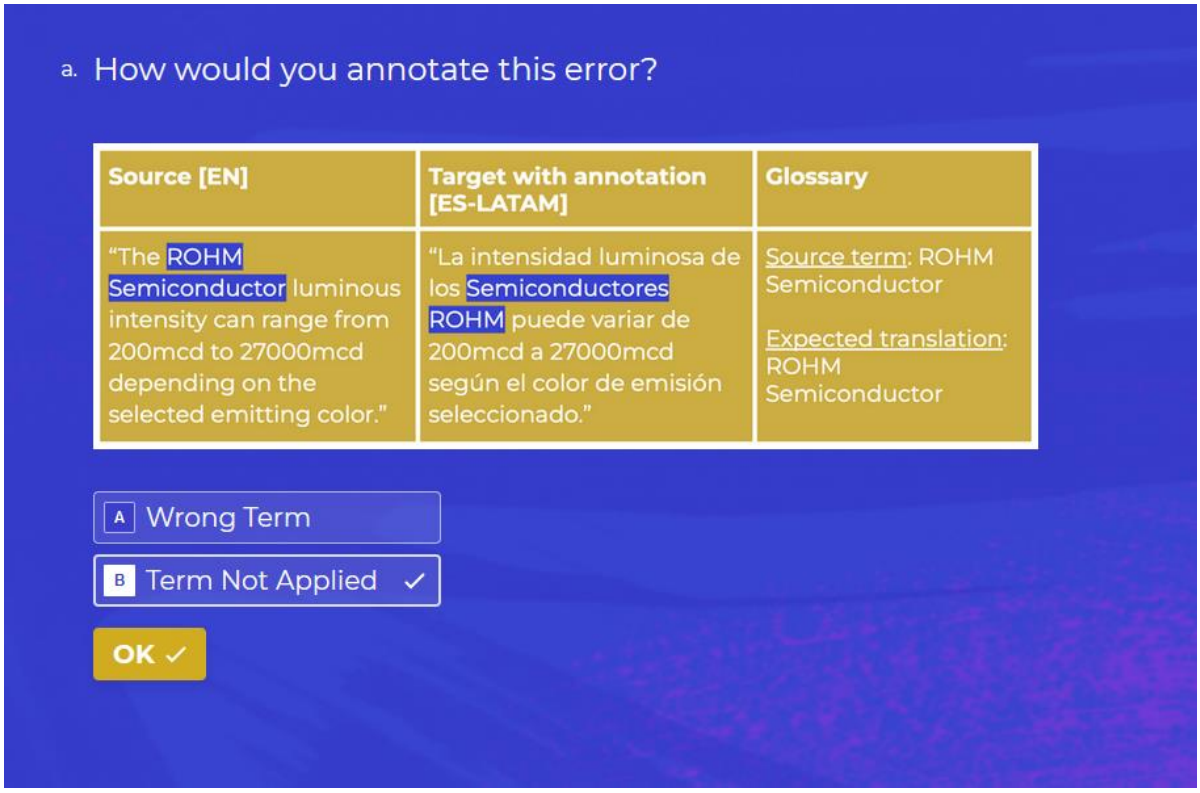

Oops...!
In this example the same principle we learned for the Whitespace tag should be used. This error falls under **Locale Conventions** and since "$79.99" is currency, the tag **Currency Format** should be used.
Continue
press Enter

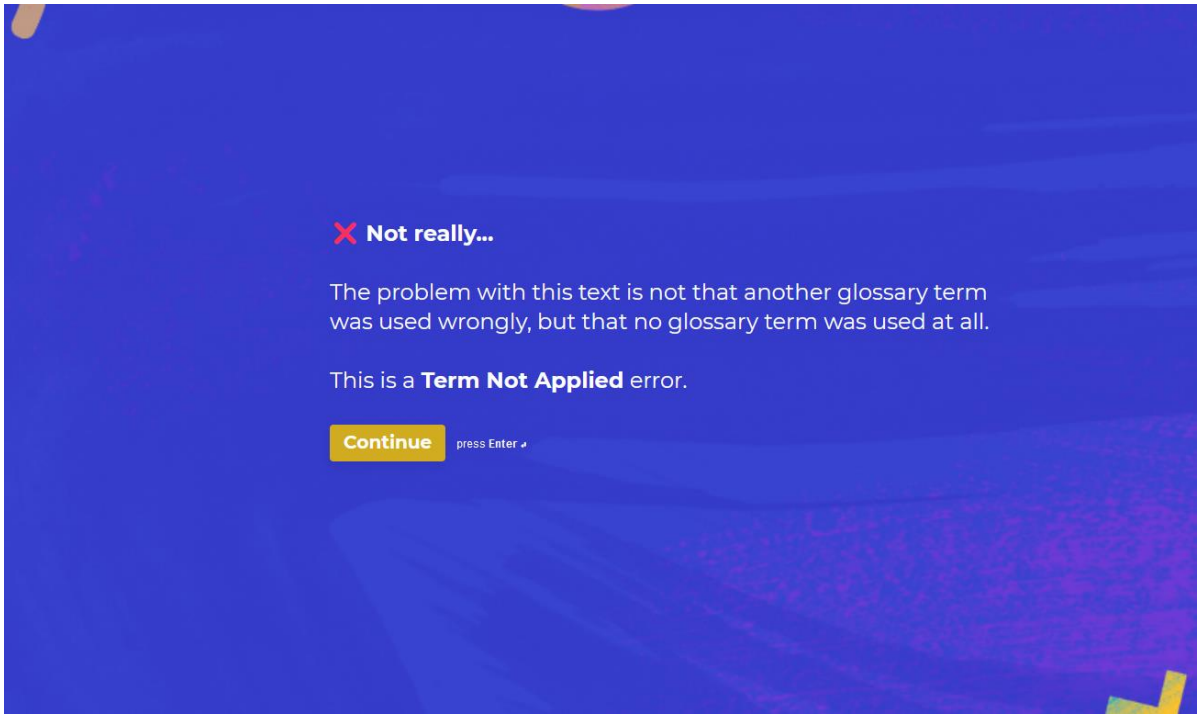

Not quite!
While there is a whitespace missing in the translation *, this error falls under **Locale Conventions** and since "$79.99" is currency, the tag **Currency Format** should be used.
*There are three different problems in this translation: the comma, the whitespace and the dollar sign in the wrong place. It should have been translated as "79,99 $".
Continue
press Enter

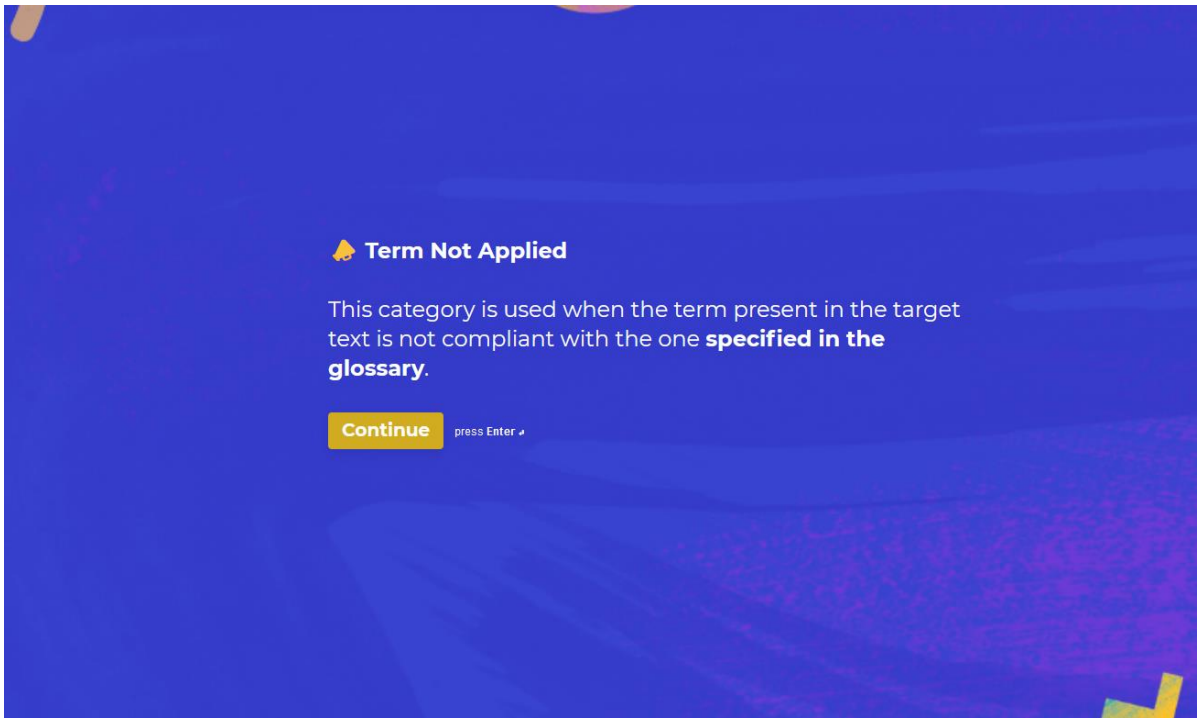

c. Would you like to see the information about Locale Conventions and Whitespaces again **to better understand your mistake**?
Y Yes
N No
OK

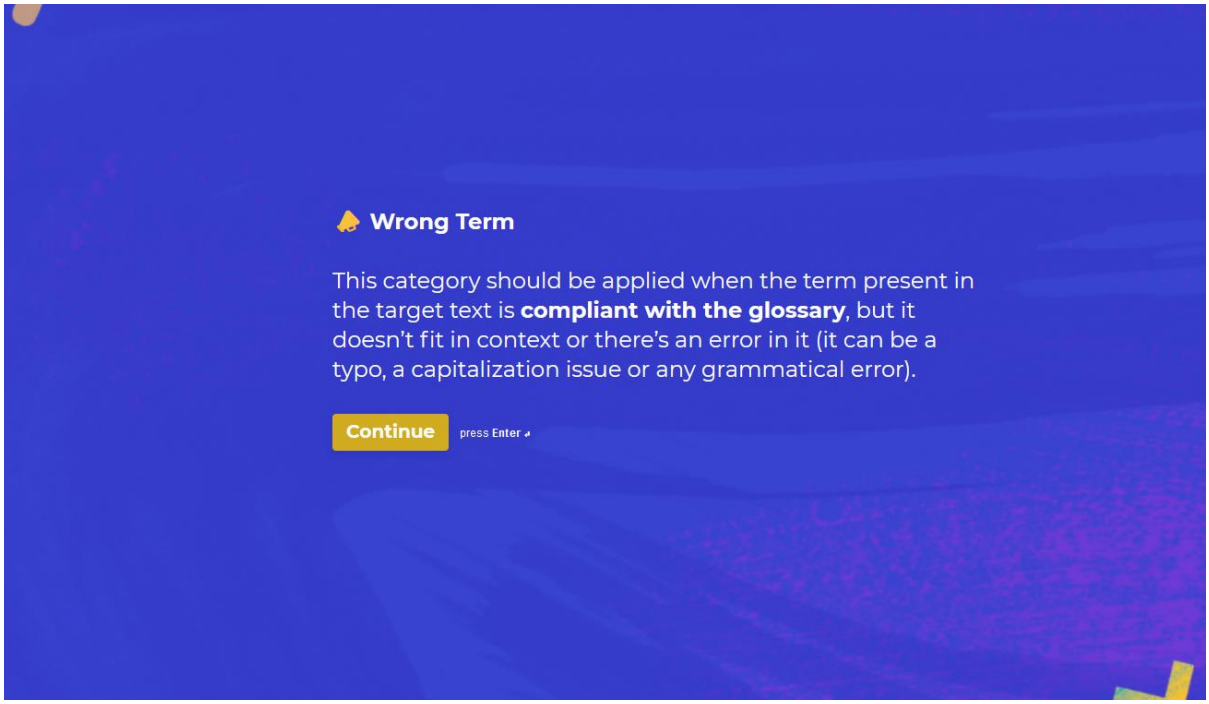

Moving on to "Terminology":
Any error that falls on any of Unbabel's glossary entries (highlighted in blue in the Annotation Tool) should be tagged with one of the Terminology subcategories.
None of the Terminology subcategories should be used to annotate anything that is not a glossary entry.
Continue
press Enter

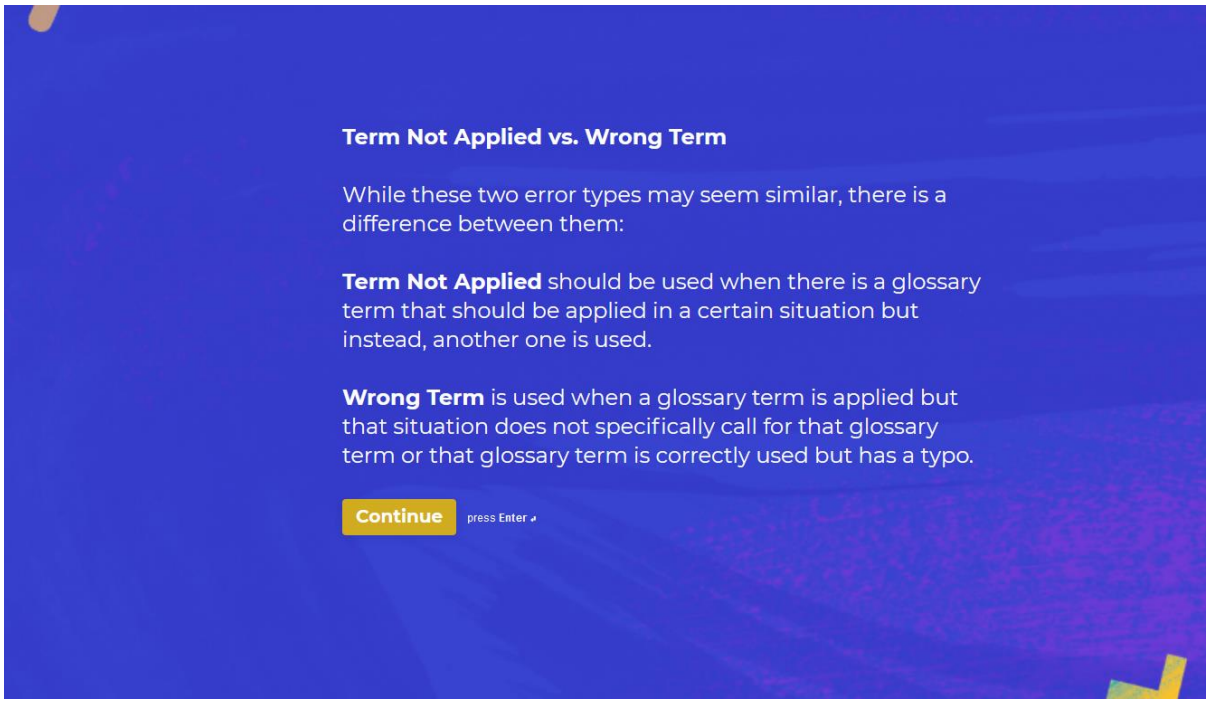

6→ Let's try some questions about Wrong Term and Term Not Applied!
Continue
press Enter


a. How would you annotate this error?
Source [EN]
Target with annotation [ES-LATAM]
Glossary
"The ROHM Semiconductor luminous intensity can range from 200mcd to 27000mcd depending on the selected emitting color."
"La intensidad luminosa de los Semiconductores ROHM puede variar de 200mcd a 27000mcd según el color de emisión seleccionado."
Source term: ROHM Semiconductor
Expected translation: ROHM Semiconductor
A Wrong Term
B Term Not Applied ✓
OK ✓


❌ Not really...
The problem with this text is not that another glossary term was used wrongly, but that no glossary term was used at all.
This is a **Term Not Applied** error.
Continue
press Enter


Term Not Applied
This category is used when the term present in the target text is not compliant with the one **specified in the glossary**.
Continue
press Enter


Wrong Term
This category should be applied when the term present in the target text is **compliant with the glossary**, but it doesn't fit in context or there's an error in it (it can be a typo, a capitalization issue or any grammatical error).
Continue
press Enter


Term Not Applied vs. Wrong Term
While these two error types may seem similar, there is a difference between them:
Term Not Applied should be used when there is a glossary term that should be applied in a certain situation but instead, another one is used.
Wrong Term is used when a glossary term is applied but that situation does not specifically call for that glossary term or that glossary term is correctly used but has a typo.
Continue
press Enter

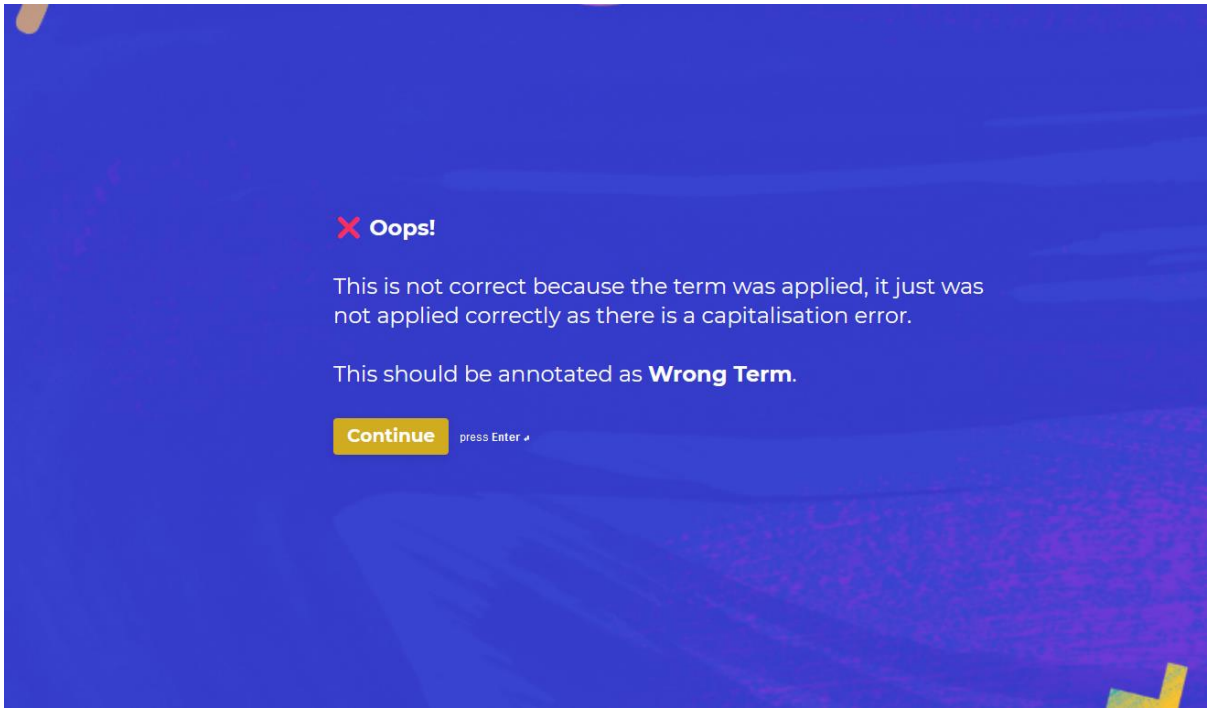
b. How would you annotate this error?

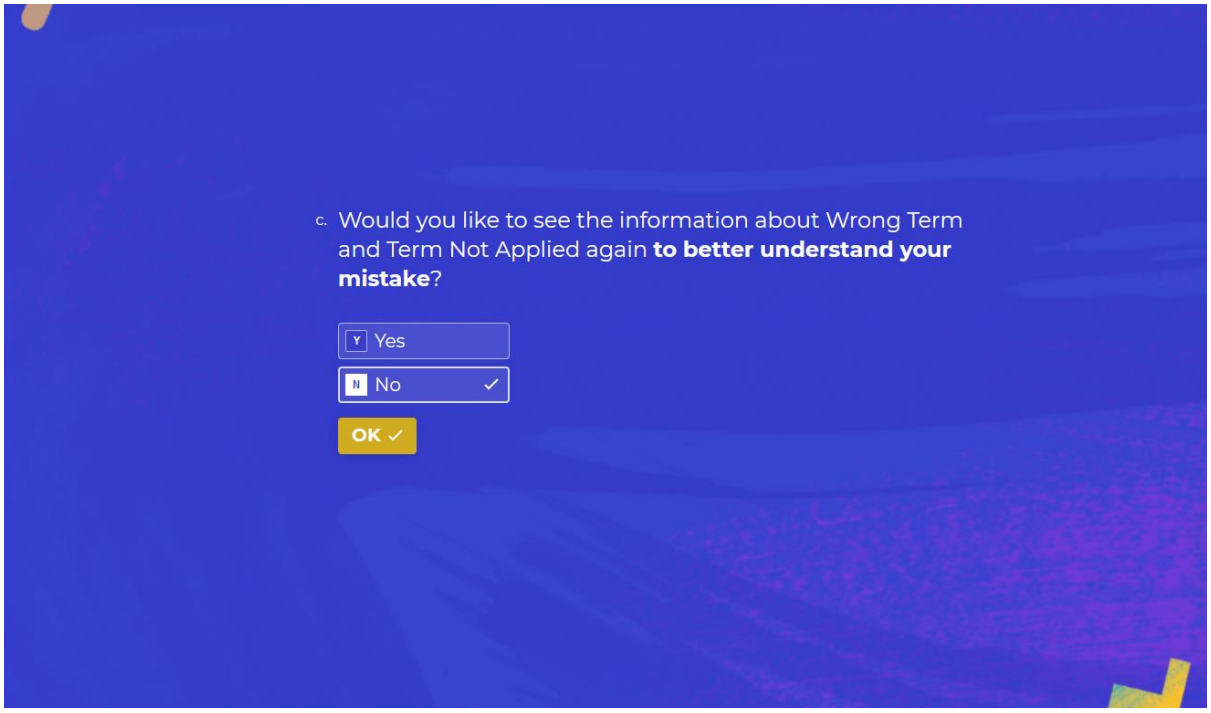
Source [EN]	Target with annotation [PT-BR]	Glossary
"Vibes will be displayed on your profile for 72 hours."	"O Vibes será exibido no seu Perfil por 72 horas."	Source term: profile Expected translation: Perfil

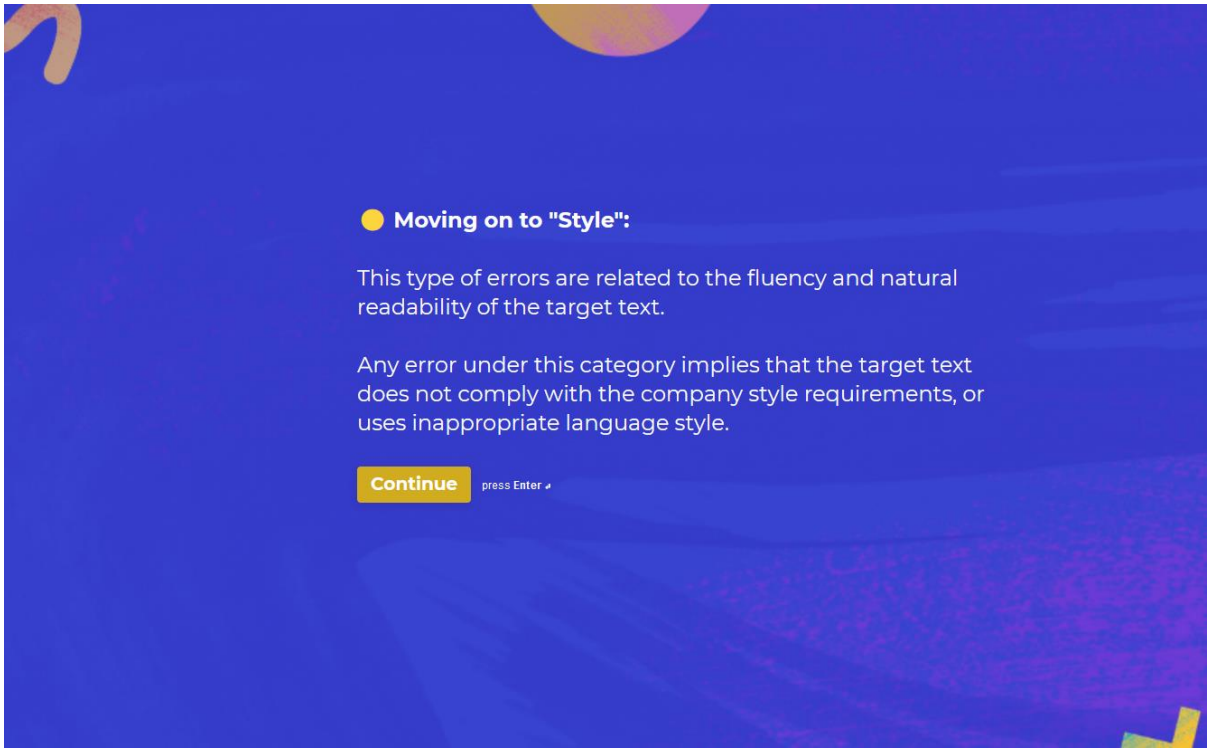
A Wrong Term ✓

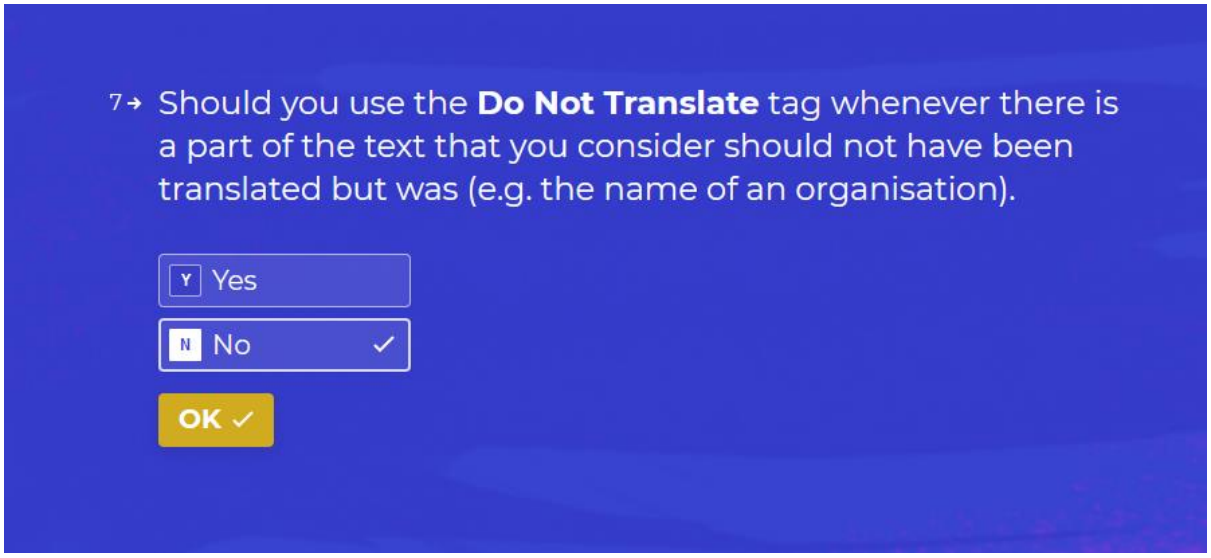
B Term Not Applied

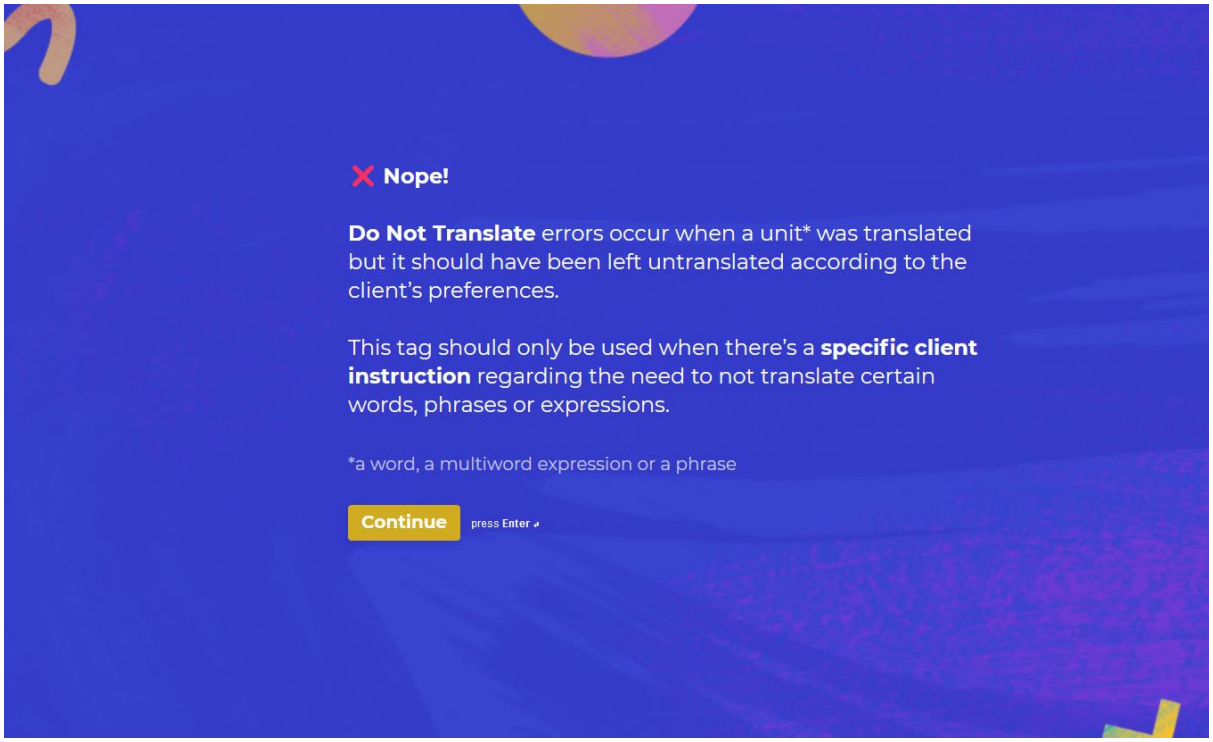
OK ✓

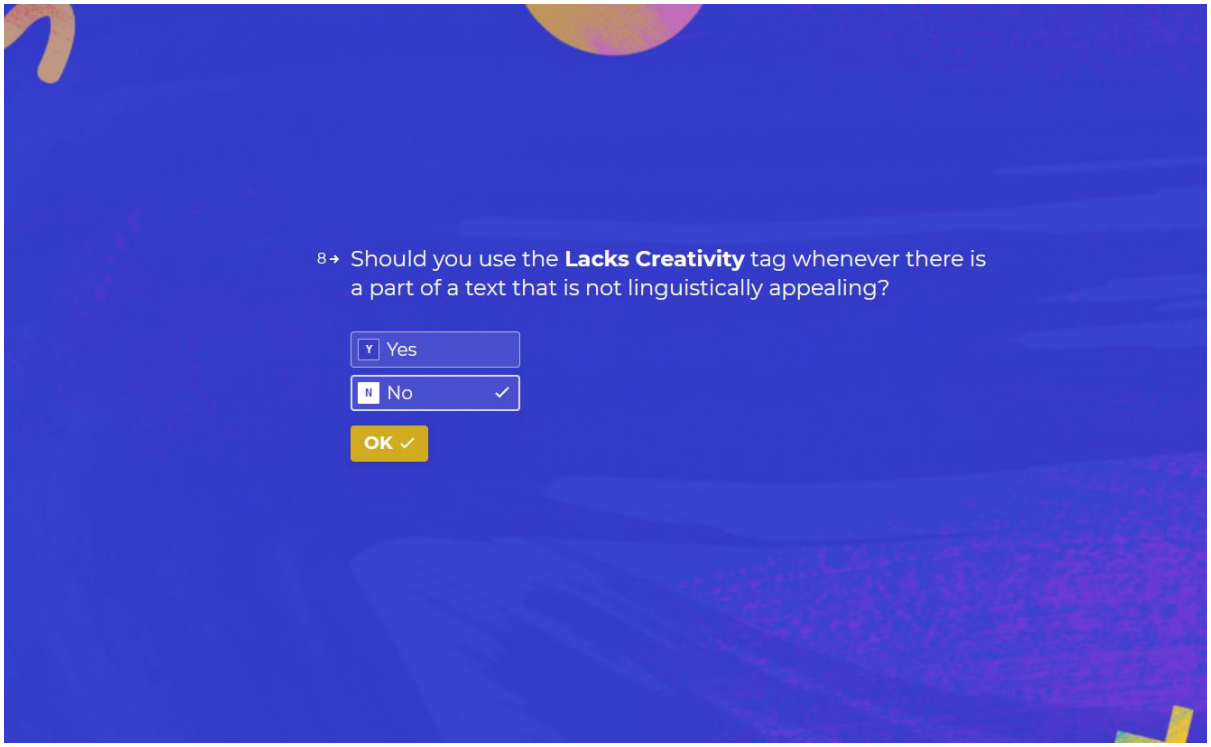


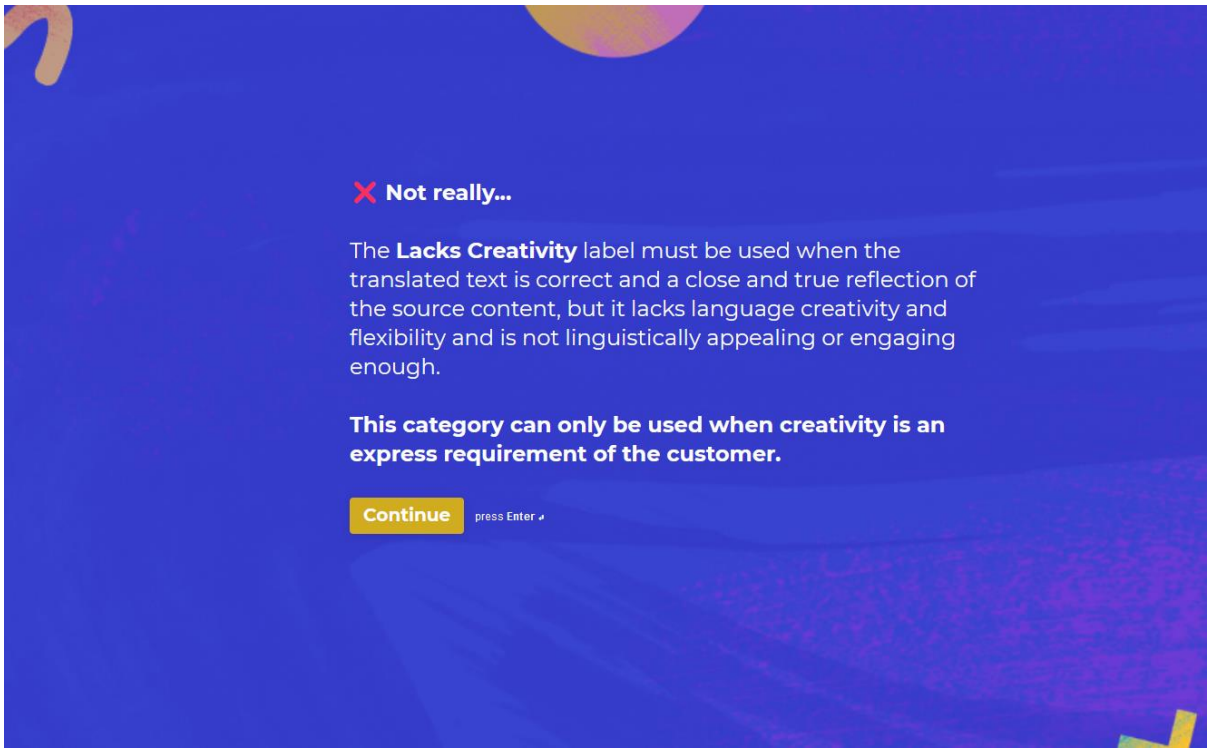

c. Would you like to see the information about Wrong Term and Term Not Applied again **to better understand your mistake**?
Y Yes
N No
OK

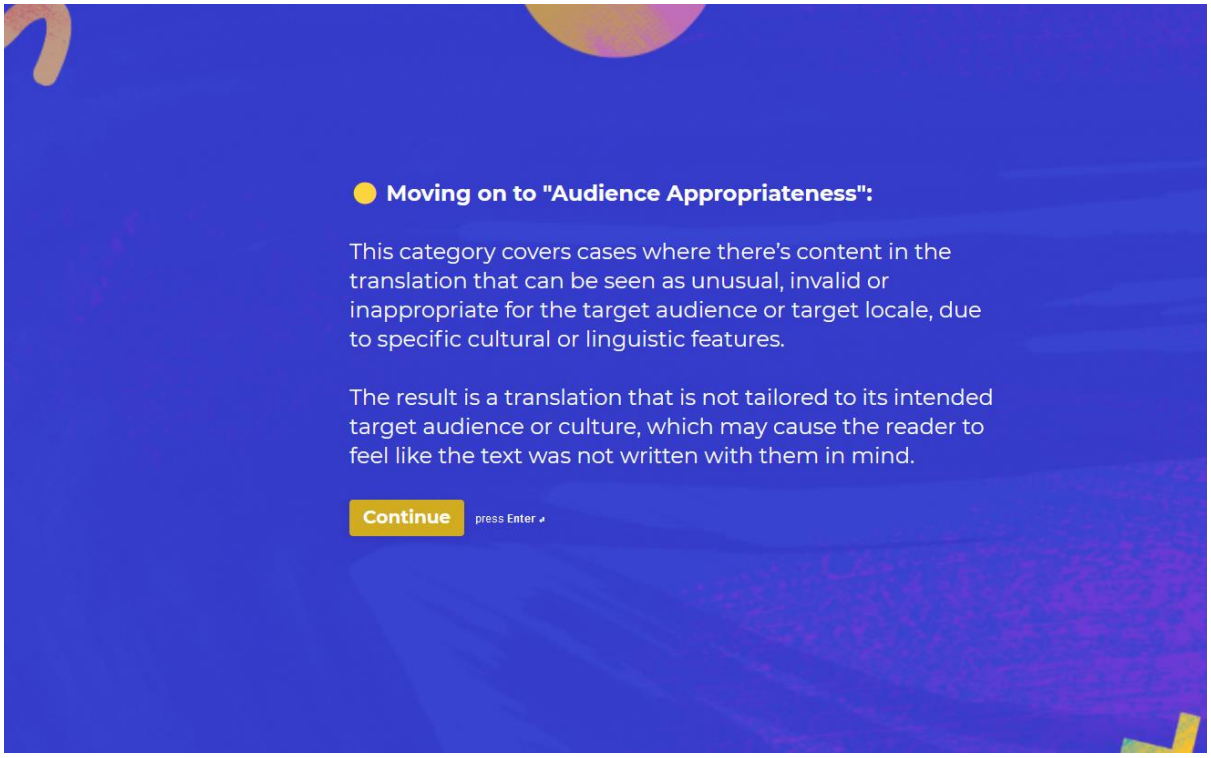

Moving on to "Style":
This type of errors are related to the fluency and natural readability of the target text.
Any error under this category implies that the target text does not comply with the company style requirements, or uses inappropriate language style.
Continue
press Enter

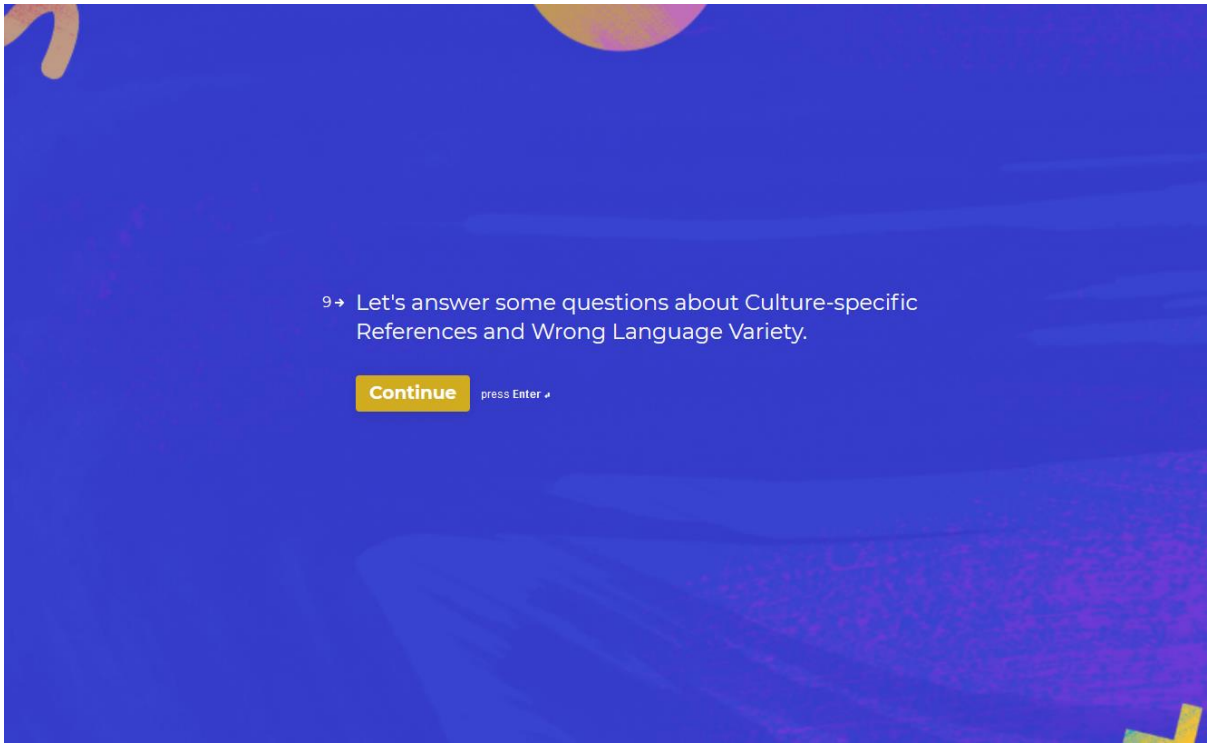

7→ Should you use the **Do Not Translate** tag whenever there is a part of the text that you consider should not have been translated but was (e.g. the name of an organisation).
Y Yes
N No ✓
OK ✓


✗ Nope!
Do Not Translate errors occur when a unit* was translated but it should have been left untranslated according to the client's preferences.
This tag should only be used when there's a **specific client instruction** regarding the need to not translate certain words, phrases or expressions.
*a word, a multiword expression or a phrase
Continue
press Enter


8→ Should you use the **Lacks Creativity** tag whenever there is a part of a text that is not linguistically appealing?
Y Yes
N No ✓
OK ✓


✗ Not really...
The **Lacks Creativity** label must be used when the translated text is correct and a close and true reflection of the source content, but it lacks language creativity and flexibility and is not linguistically appealing or engaging enough.
This category can only be used when creativity is an express requirement of the customer.
Continue
press Enter


Moving on to "Audience Appropriateness":
This category covers cases where there's content in the translation that can be seen as unusual, invalid or inappropriate for the target audience or target locale, due to specific cultural or linguistic features.
The result is a translation that is not tailored to its intended target audience or culture, which may cause the reader to feel like the text was not written with them in mind.
Continue
press Enter


9→ Let's answer some questions about Culture-specific References and Wrong Language Variety.
Continue
press Enter

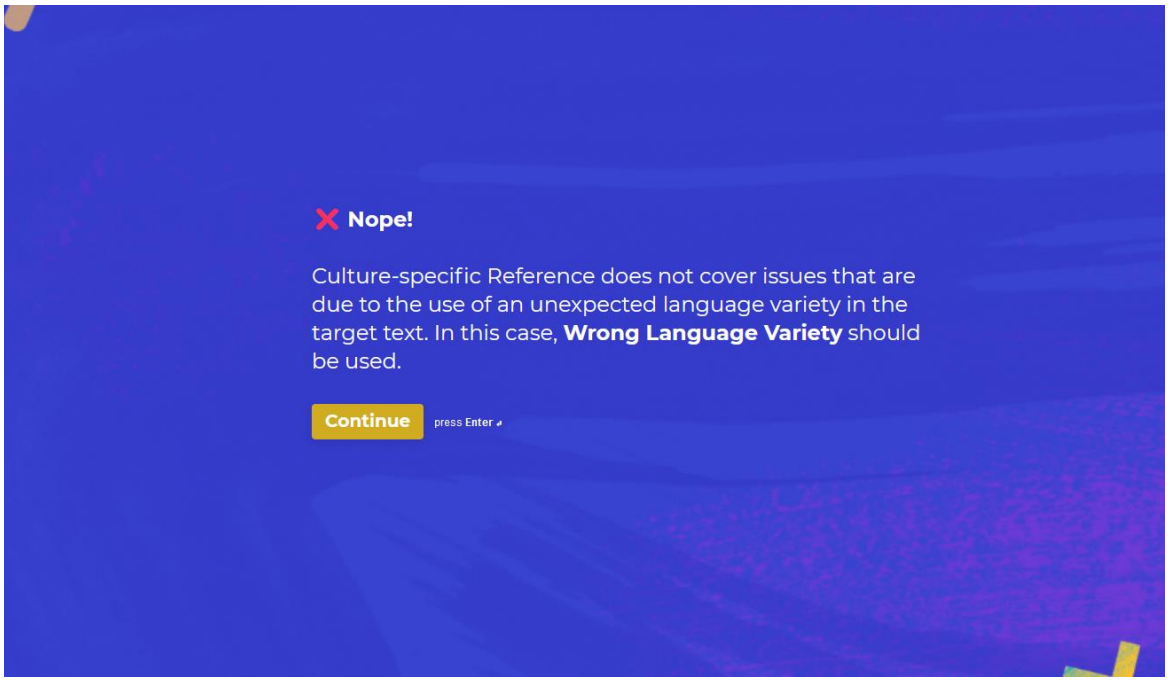
a. How would you annotate this error?

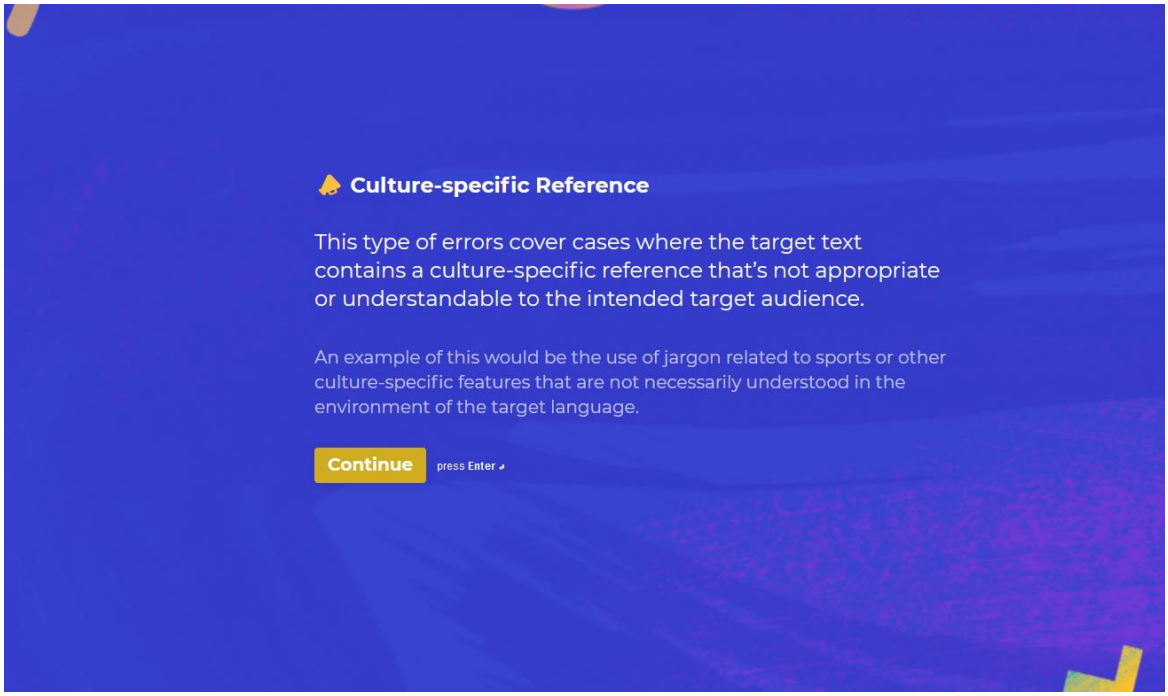
Source [EN]	Target with annotation [PT-PT]	Context
"Stephen will be instrumental in strengthening and leading the EMEA teams."	"Stephen será fundamental para fortalecer e liderar as equipes da EMEA."	The term "equipes" is Brazilian Portuguese and not European Portuguese.

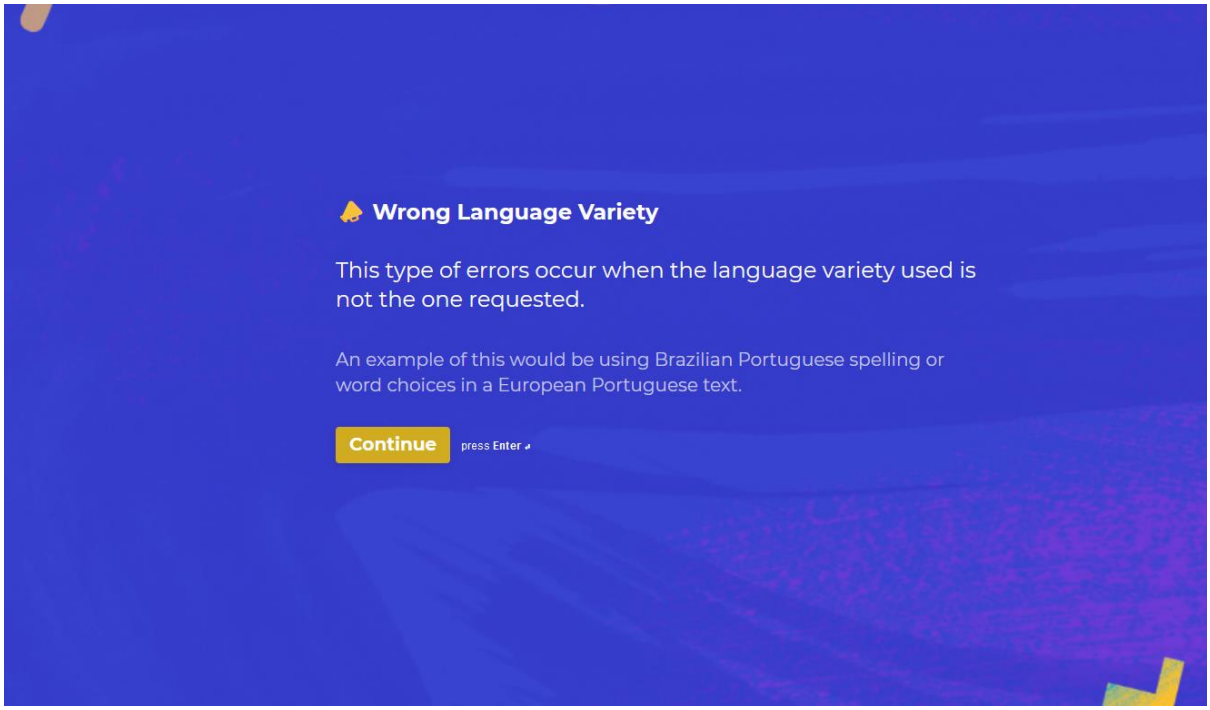
A Culture-specific Reference

B Wrong Language Variety ✓

OK ✓




Culture-specific Reference
This type of errors cover cases where the target text contains a culture-specific reference that's not appropriate or understandable to the intended target audience.
An example of this would be the use of jargon related to sports or other culture-specific features that are not necessarily understood in the environment of the target language.
Continue
press Enter


Wrong Language Variety
This type of errors occur when the language variety used is not the one requested.
An example of this would be using Brazilian Portuguese spelling or word choices in a European Portuguese text.
Continue
press Enter

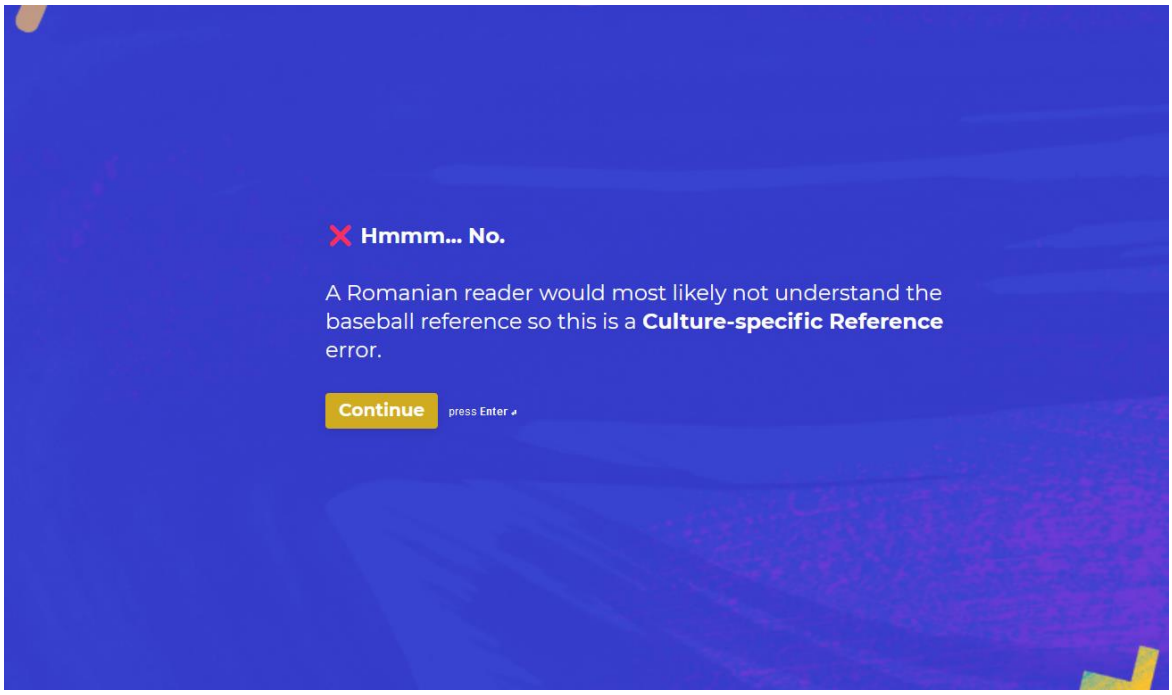
b. How would you annotate this error?

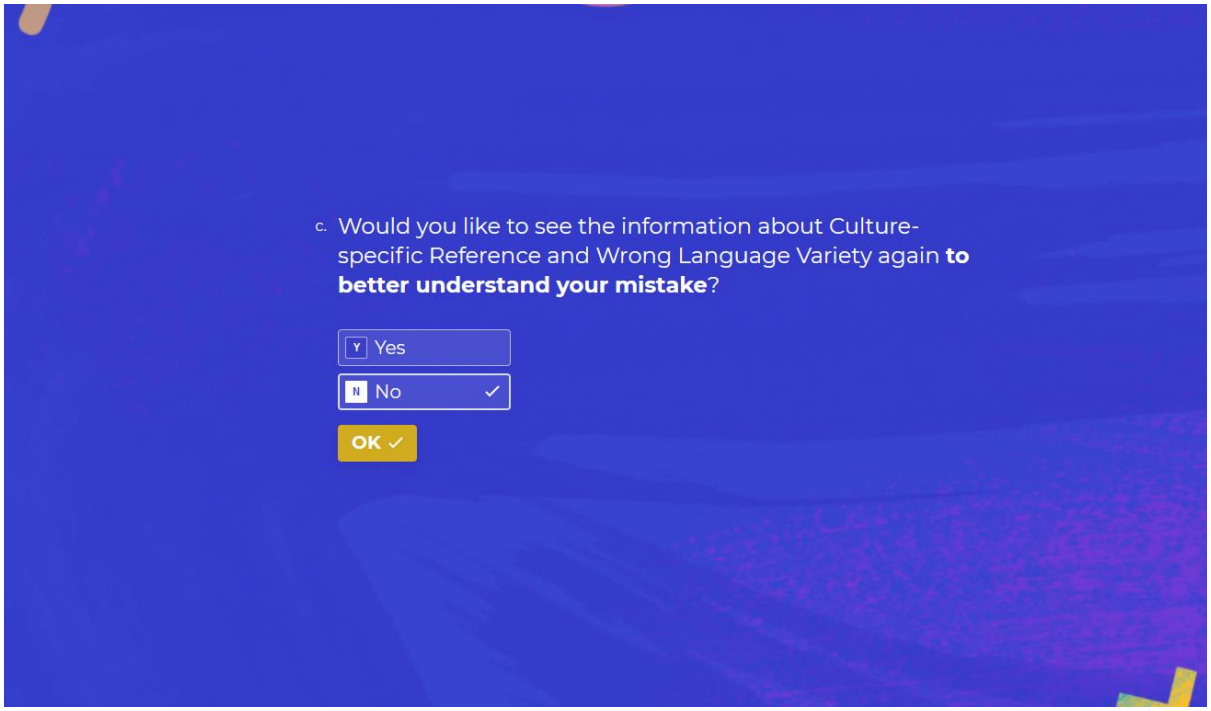
Source [EN]	Target with annotation [RO]	Context
"The president's speech was a home run."	"Discursul președintelui a fost un home run."	The source text uses a metaphor from baseball, which would not make any sense to a Romanian audience.

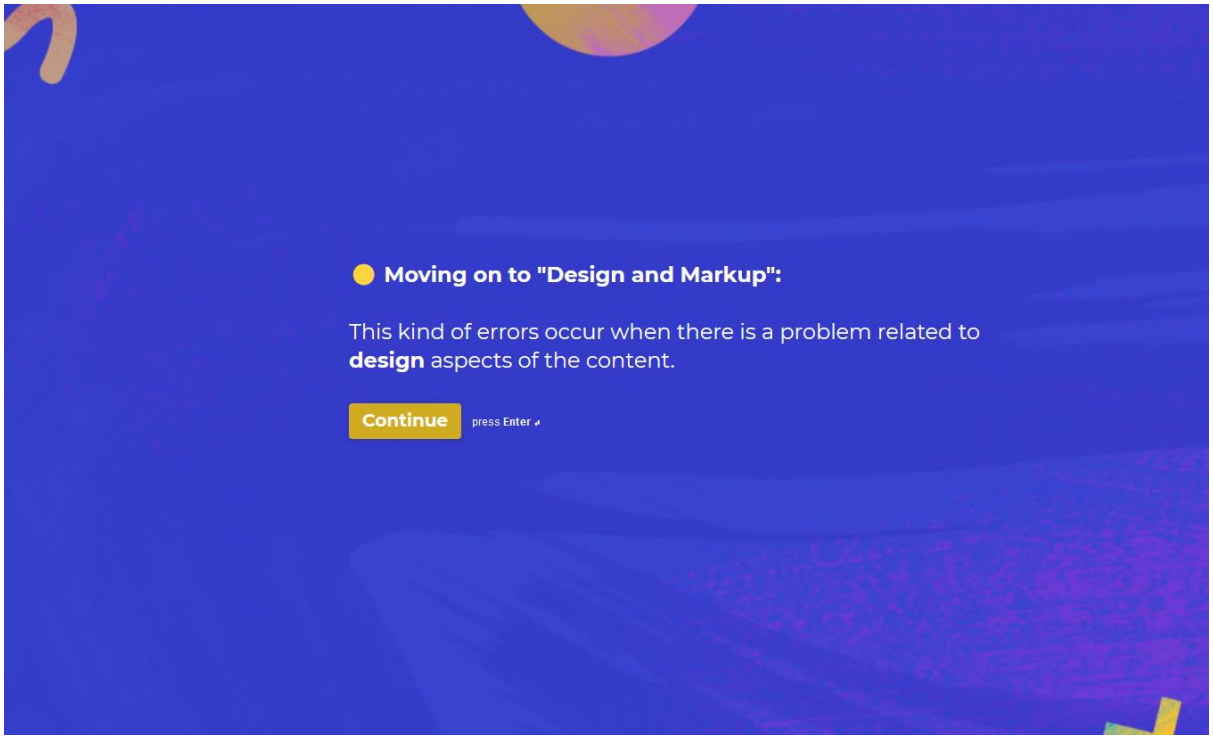
A Culture-specific Reference ✓

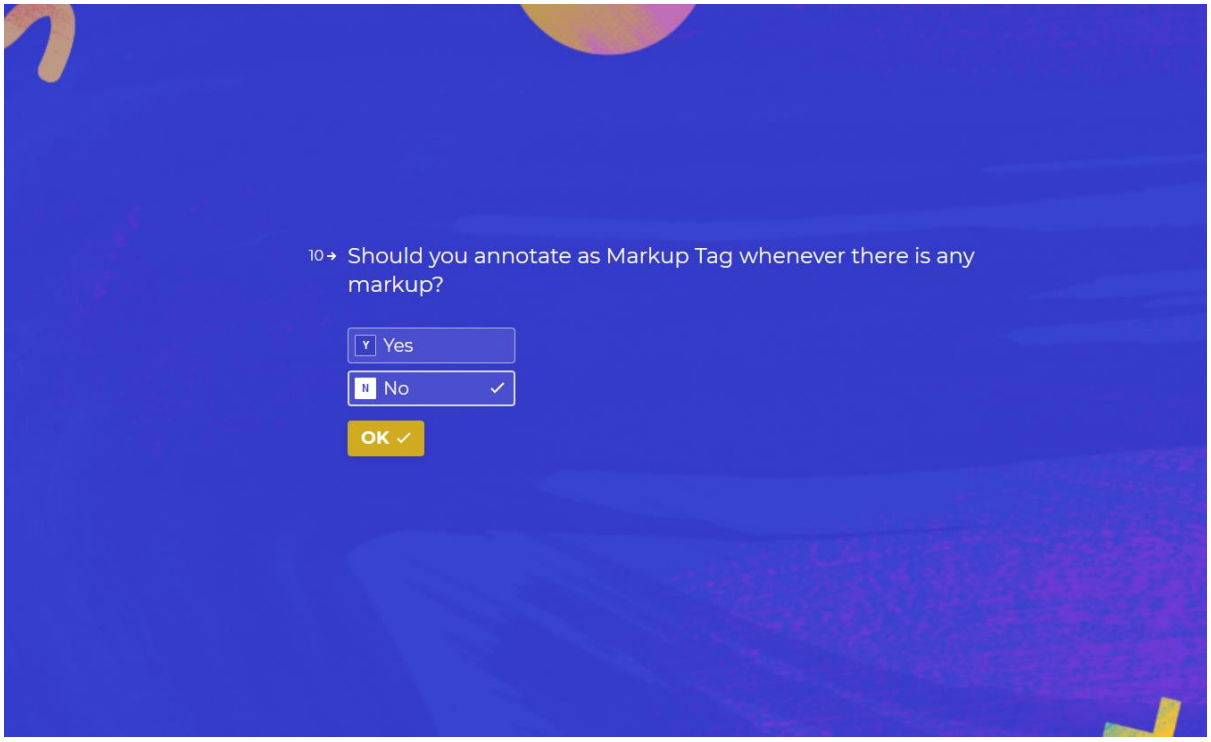
B Wrong Language Variety

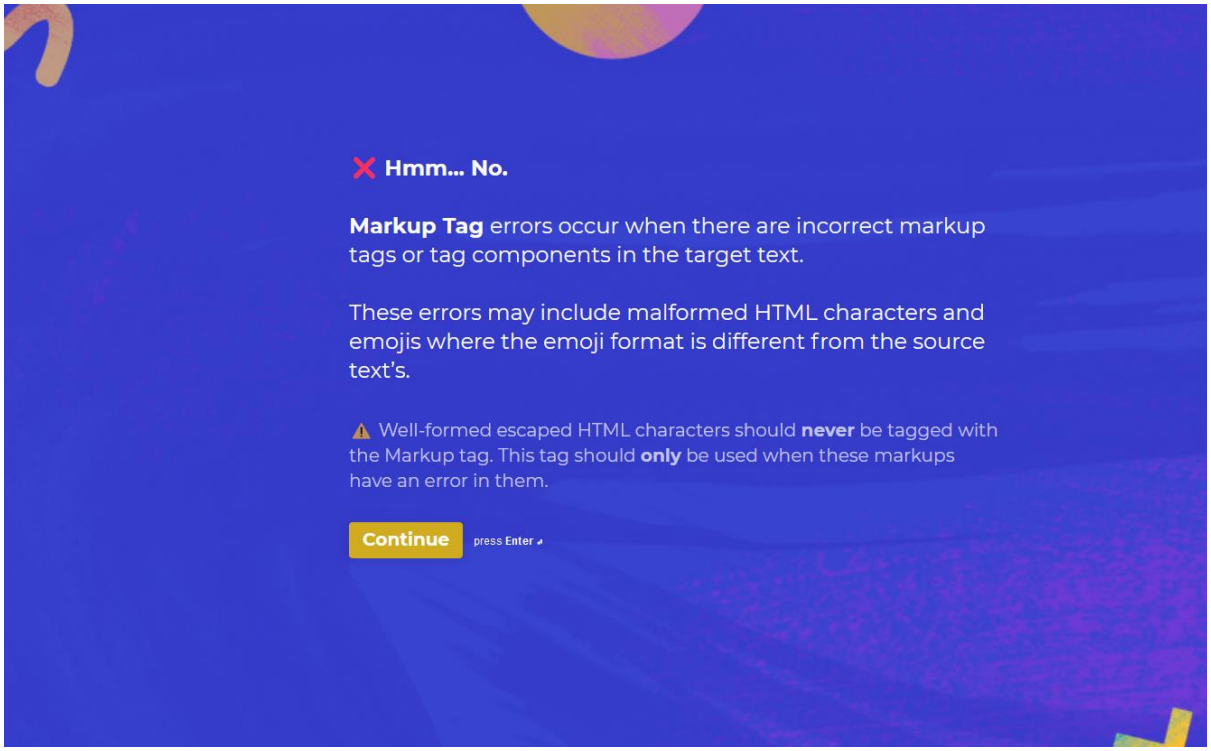
OK ✓




c. Would you like to see the information about Culture-specific Reference and Wrong Language Variety again **to better understand your mistake**?
Y Yes
N No
OK


Moving on to "Design and Markup":
This kind of errors occur when there is a problem related to **design** aspects of the content.
Continue
press Enter


10 → Should you annotate as Markup Tag whenever there is any markup?
Y Yes
N No ✓
OK ✓


❌ Hmm... No.
Markup Tag errors occur when there are incorrect markup tags or tag components in the target text.
These errors may include malformed HTML characters and emojis where the emoji format is different from the source text's.
⚠ Well-formed escaped HTML characters should never be tagged with the Markup tag. This tag should only be used when these markups have an error in them.
Continue
press Enter

The following are examples of the most common, correctly escaped HTML characters in our data:

⚠ These are not errors and shouldn't be annotated

Continue press Enter

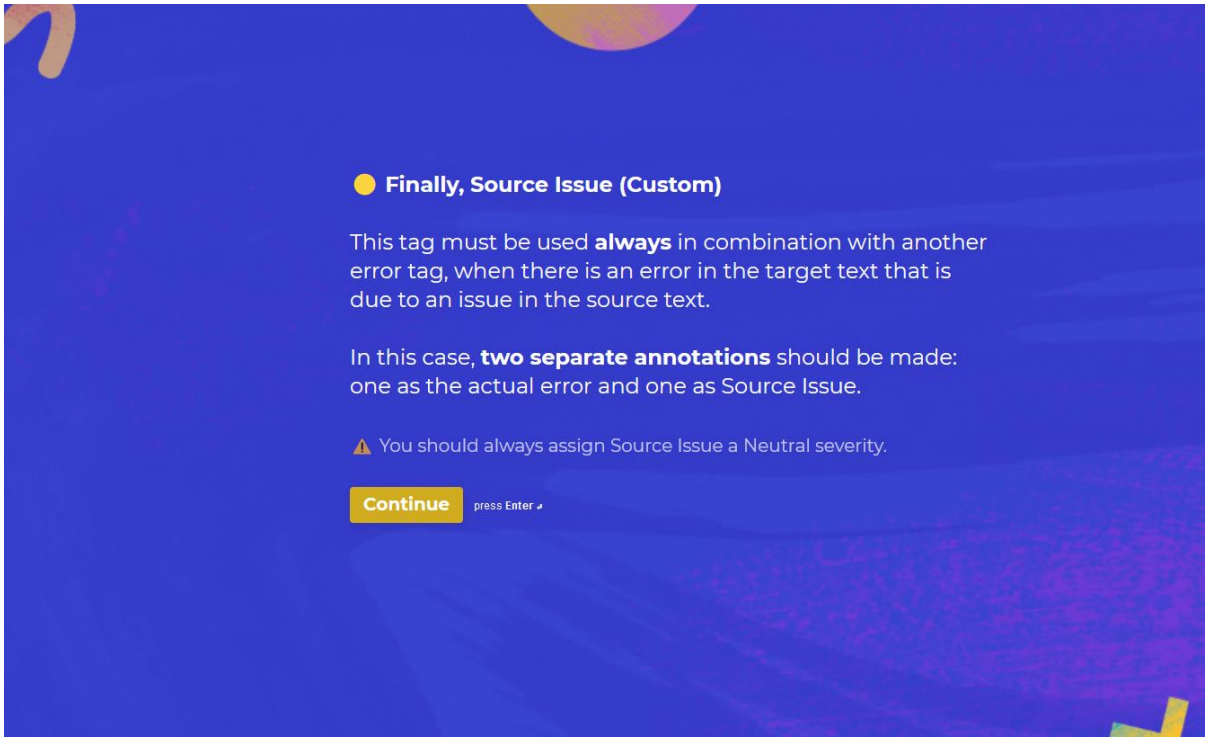
Well-formed HTML (Do not annotate)	What it stands for
&#91;	[
&#93;	[
&#124;	|
< or <	<
> or >	>
" or "	"
' or '	'
& or &	&

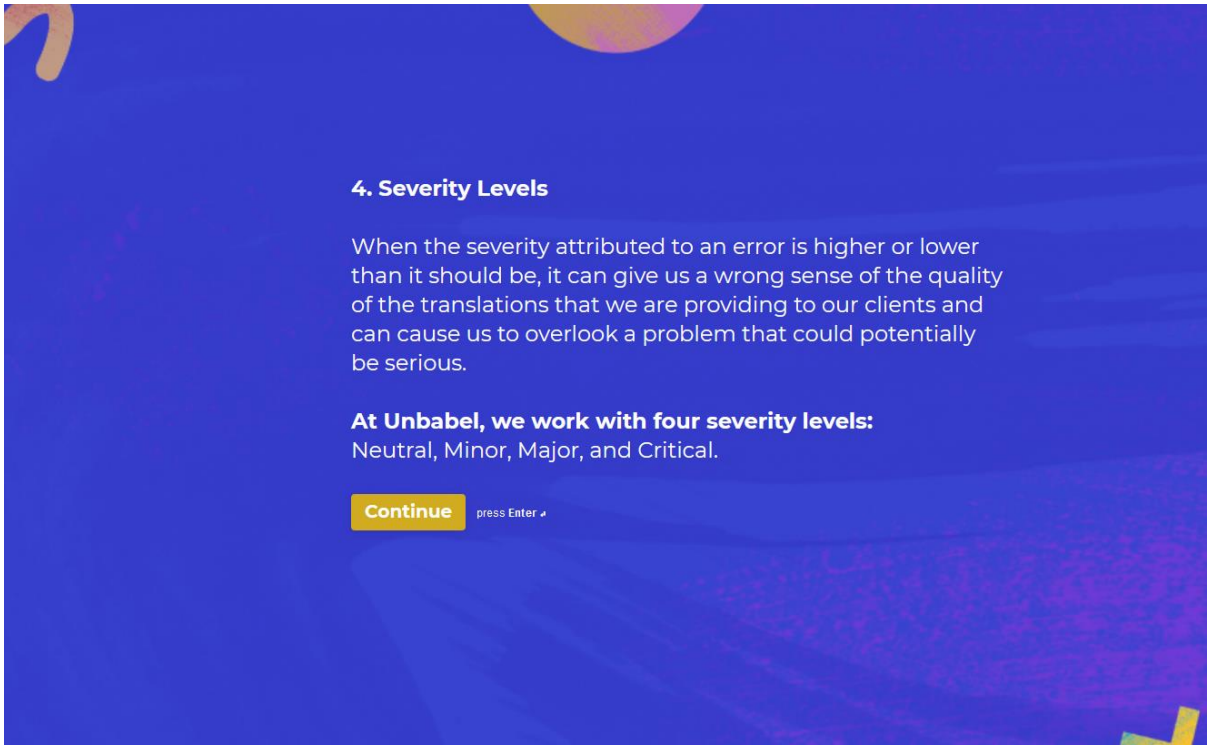
These are examples of malformed HTML codes that should be annotated:

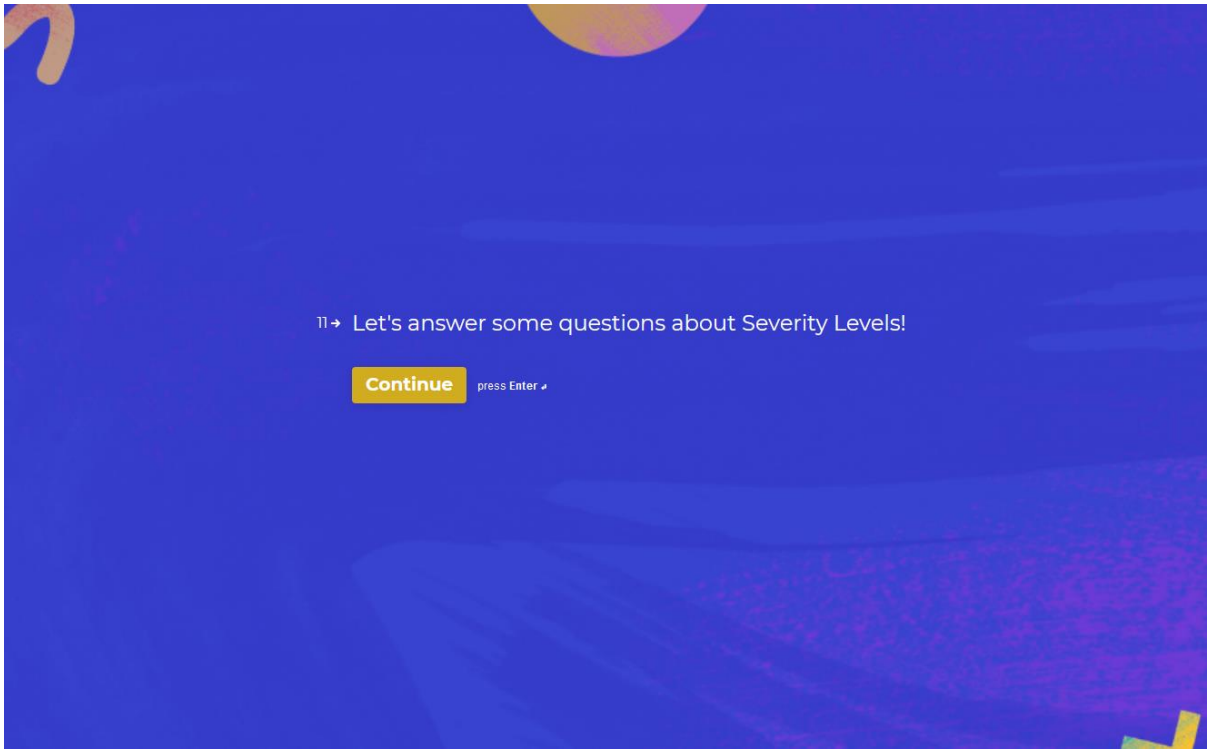
- & apos;
- & APOS ;
- & AMP;
- > ;

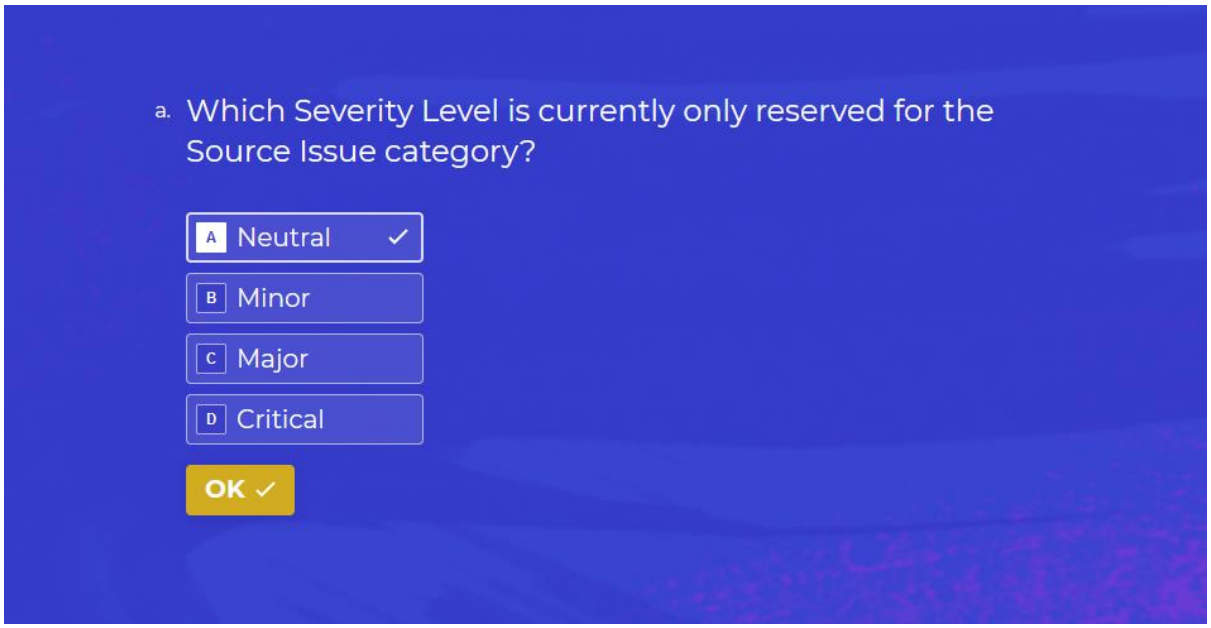
As you can see, these markups are incorrect because, unlike the ones we saw previously, they have spaces that should not be there.

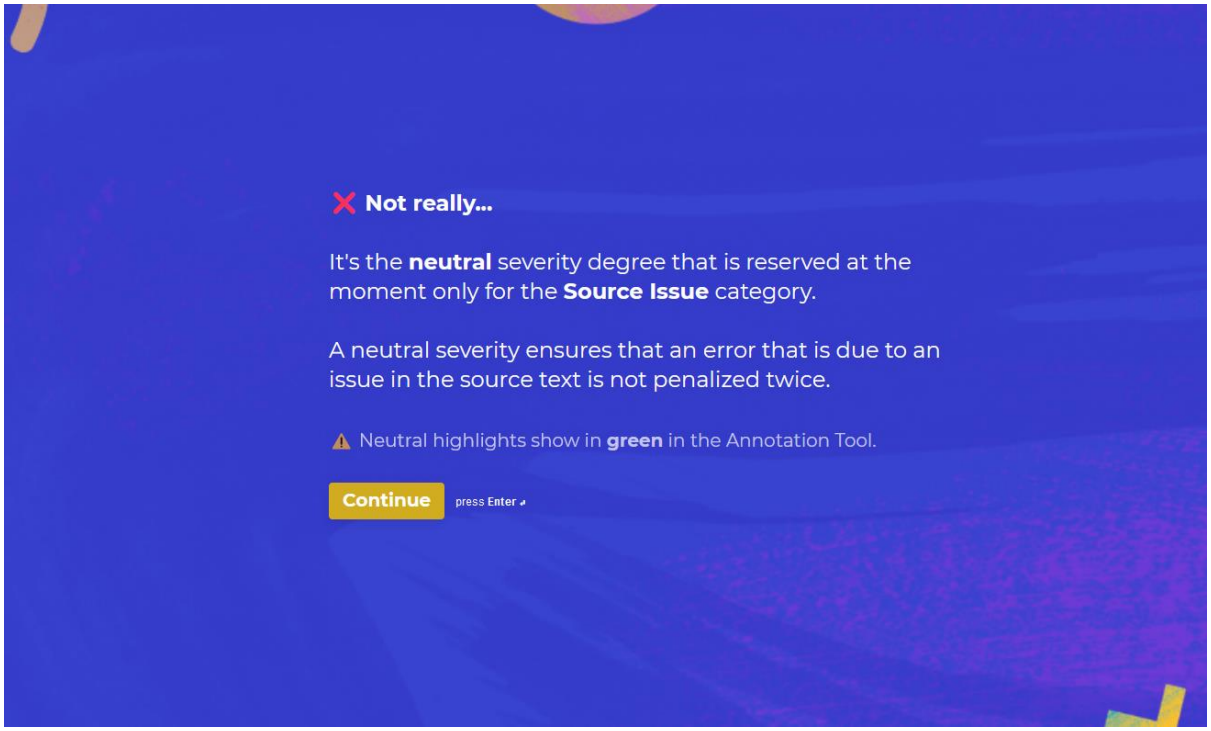
Continue press Enter

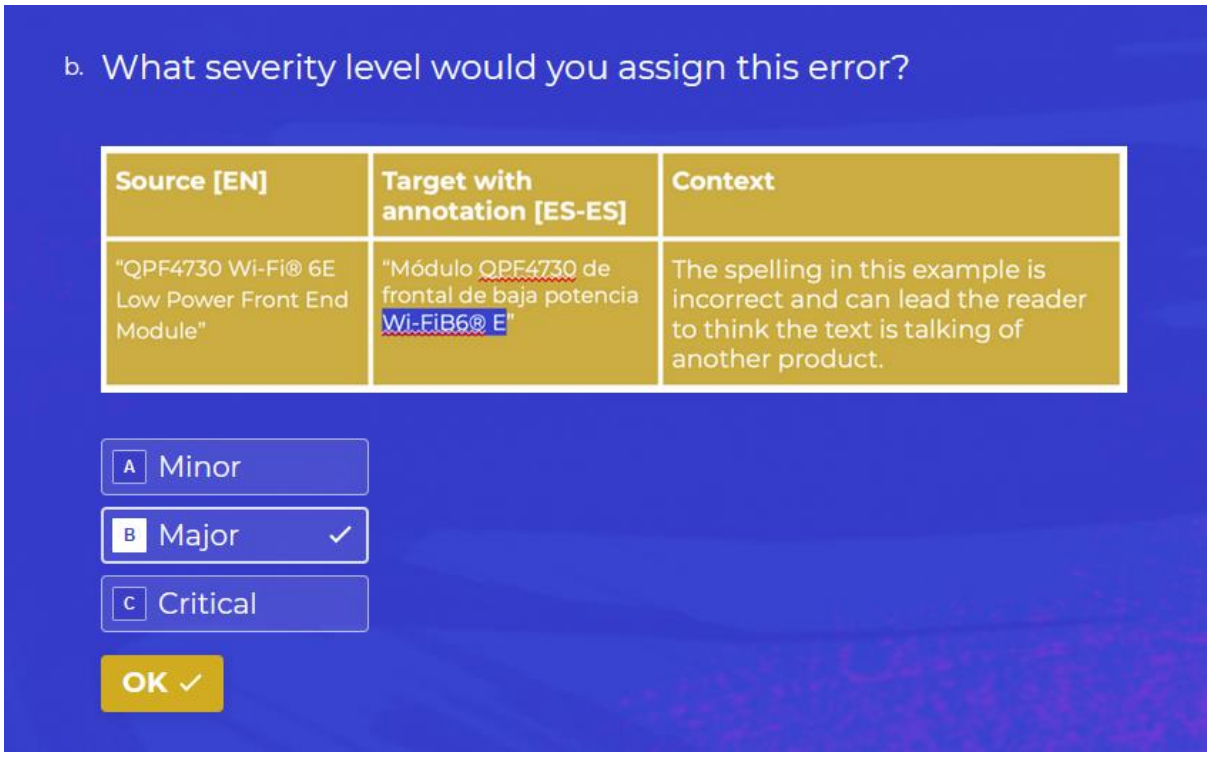

Finally, Source Issue (Custom)
This tag must be used always in combination with another error tag, when there is an error in the target text that is due to an issue in the source text.
In this case, two separate annotations should be made: one as the actual error and one as Source Issue.
You should always assign Source Issue a Neutral severity.
Continue
press Enter

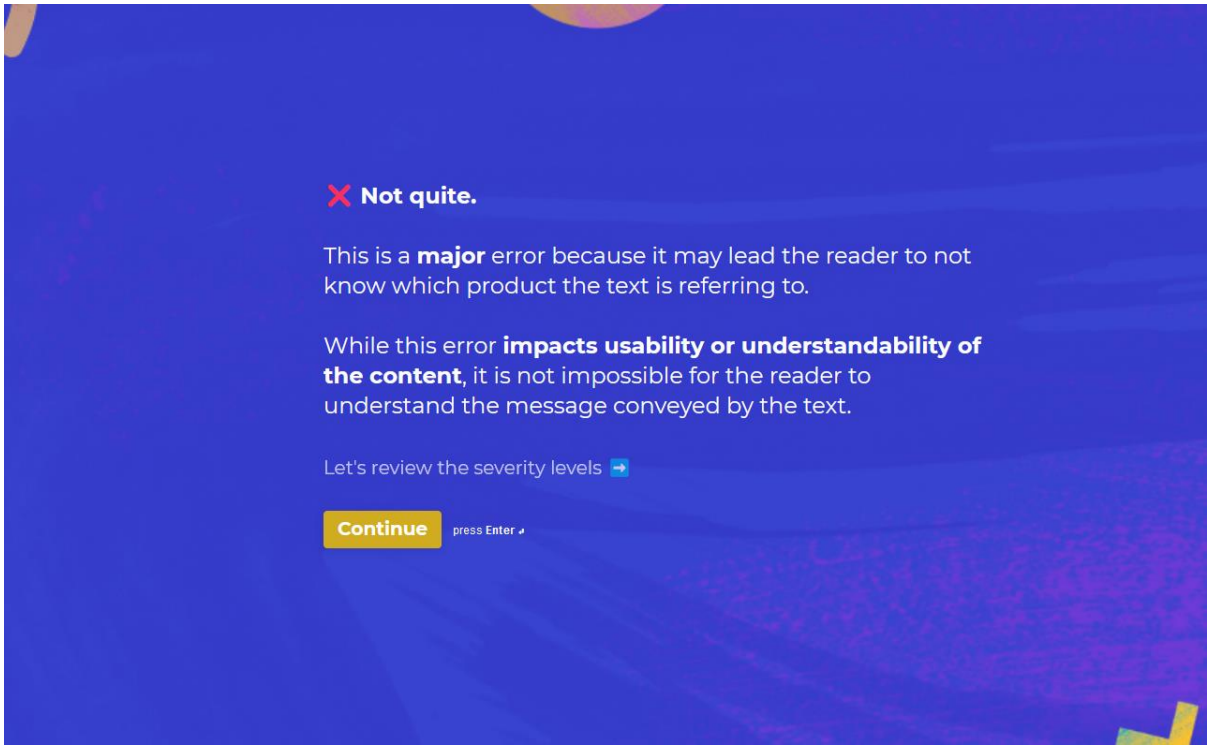

4. Severity Levels
When the severity attributed to an error is higher or lower than it should be, it can give us a wrong sense of the quality of the translations that we are providing to our clients and can cause us to overlook a problem that could potentially be serious.
At Unbabel, we work with four severity levels:
Neutral, Minor, Major, and Critical.
Continue
press Enter

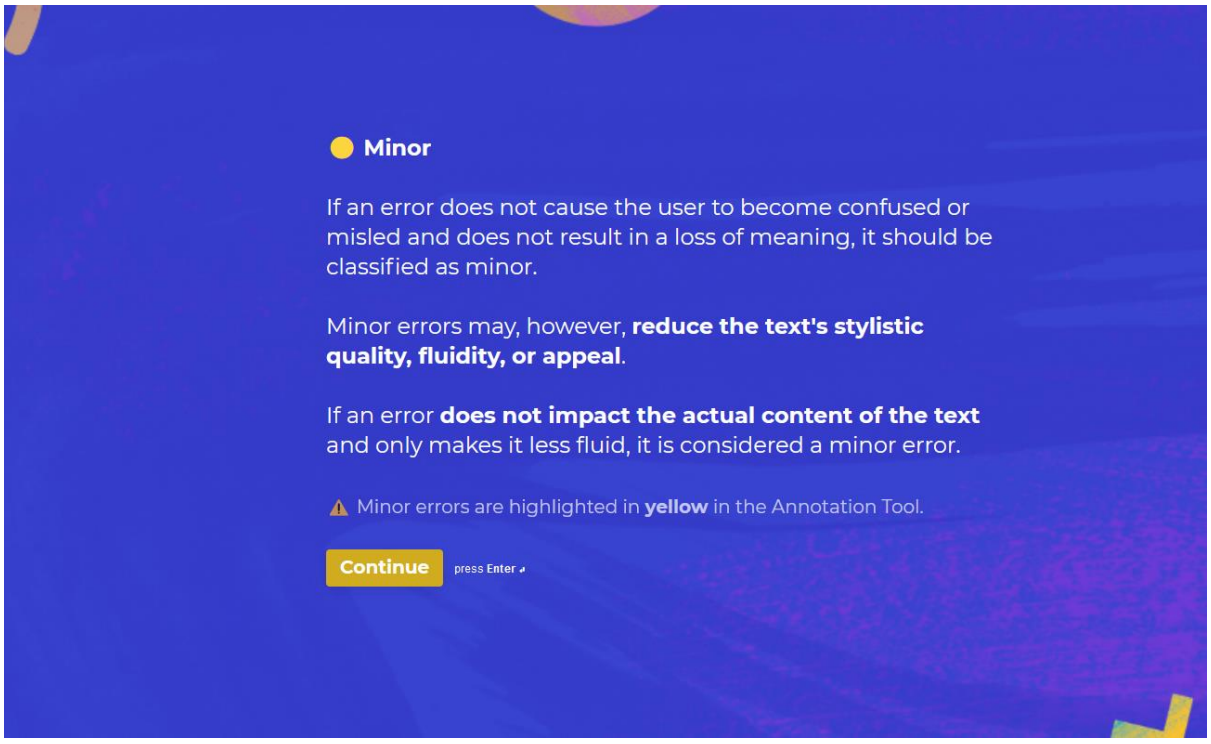

Let's answer some questions about Severity Levels!
Continue
press Enter

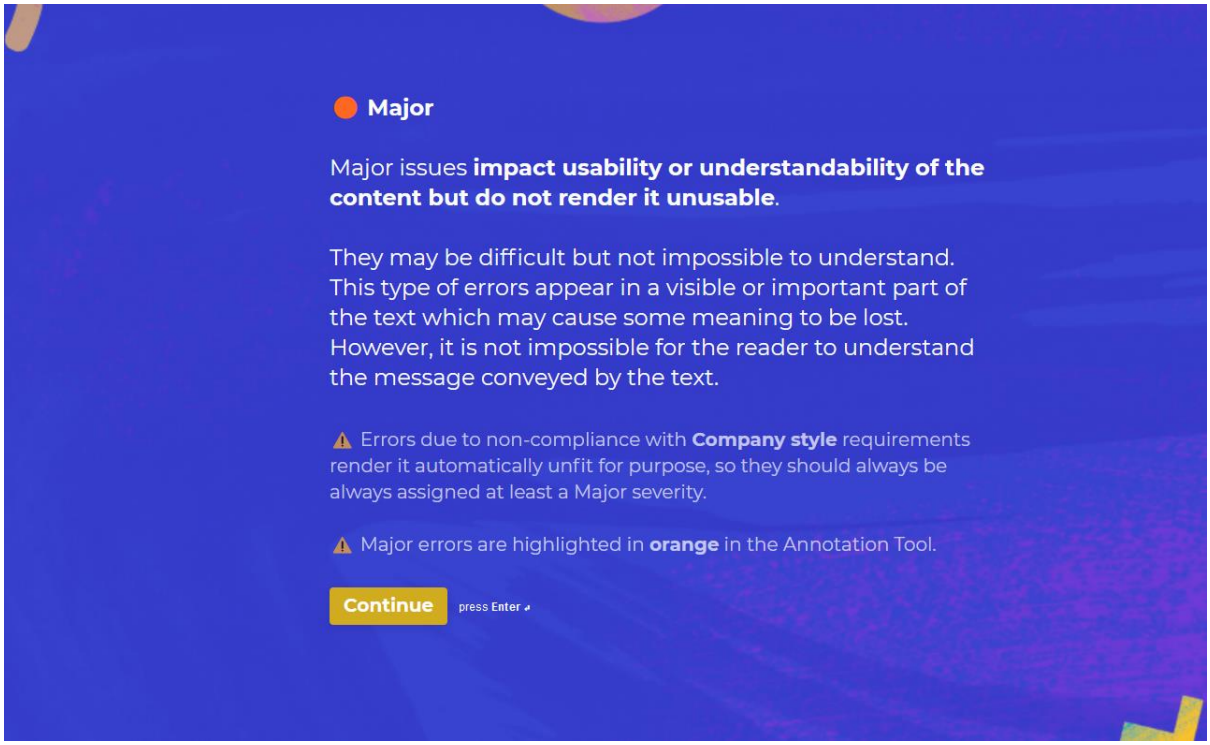

a. Which Severity Level is currently only reserved for the Source Issue category?
A Neutral
B Minor
C Major
D Critical
OK

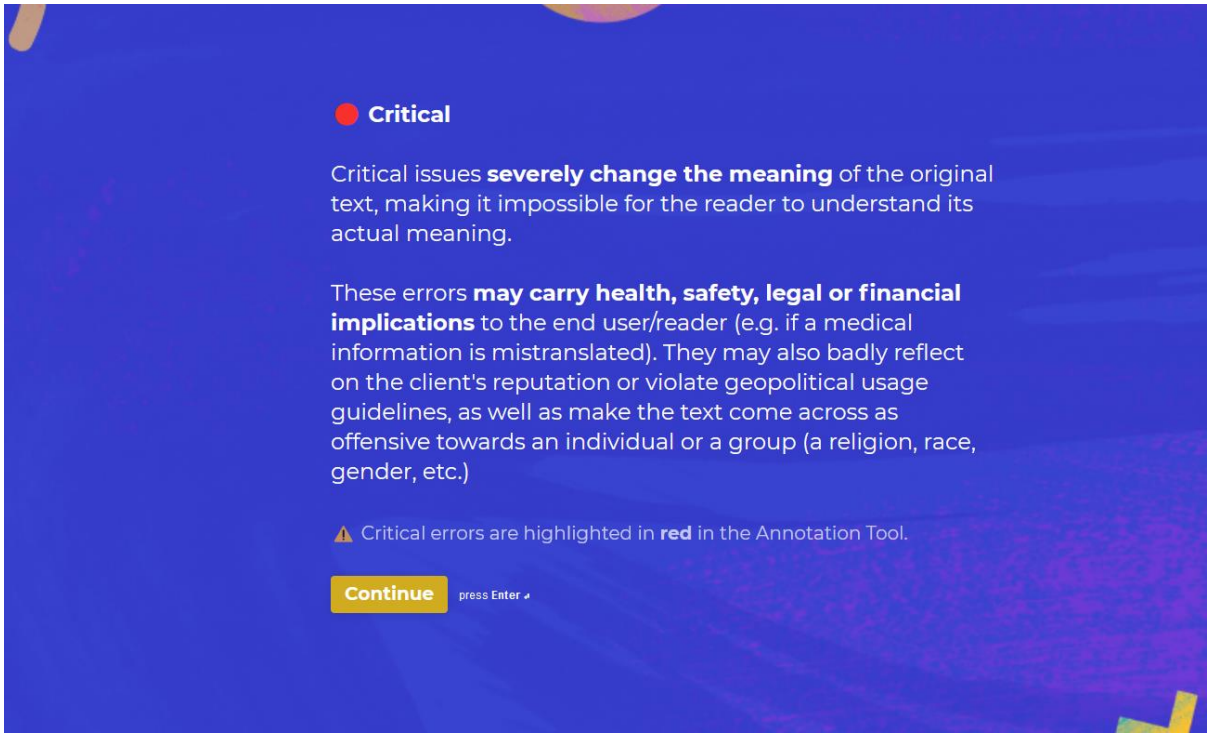

Not really...
It's the neutral severity degree that is reserved at the moment only for the Source Issue category.
A neutral severity ensures that an error that is due to an issue in the source text is not penalized twice.
Neutral highlights show in green in the Annotation Tool.
Continue
press Enter

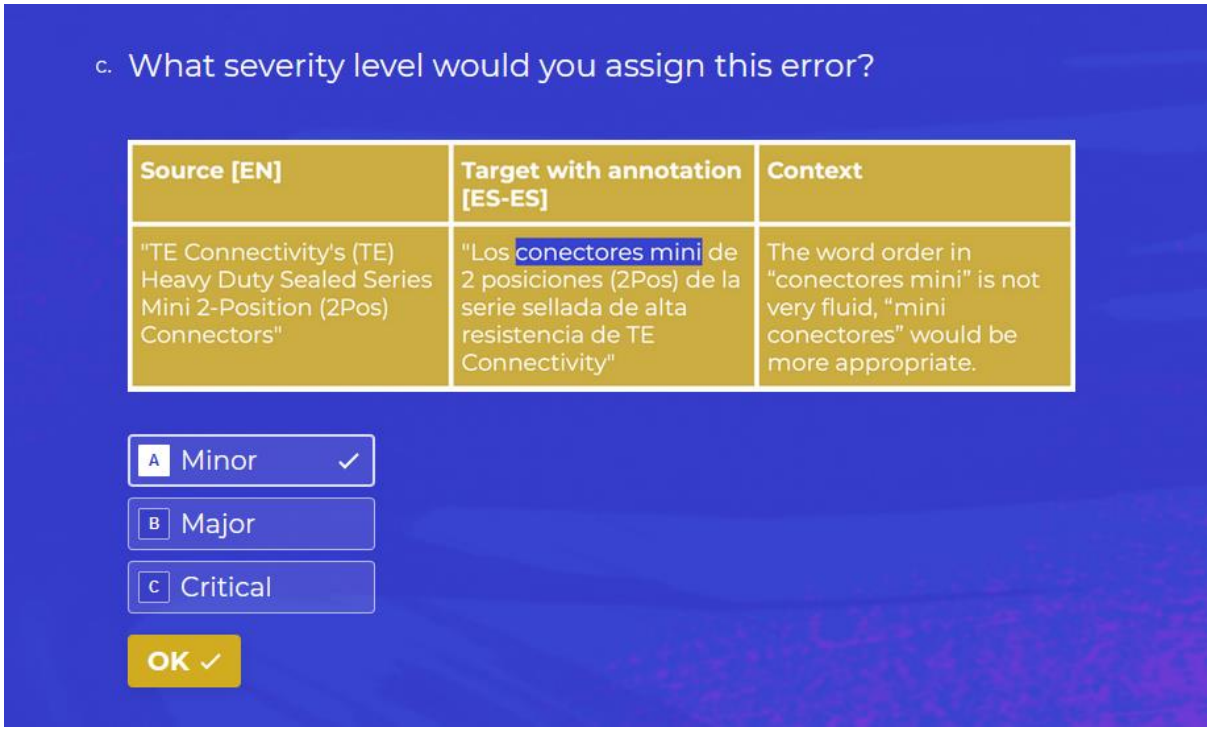

b. What severity level would you assign this error?
Source [EN]
Target with annotation [ES-ES]
Context
"QPF4730 Wi-Fi® 6E Low Power Front End Module"
"Módulo QPF4730 de frontal de baja potencia Wi-FiB6® E"
The spelling in this example is incorrect and can lead the reader to think the text is talking of another product.
A Minor
B Major
C Critical
OK

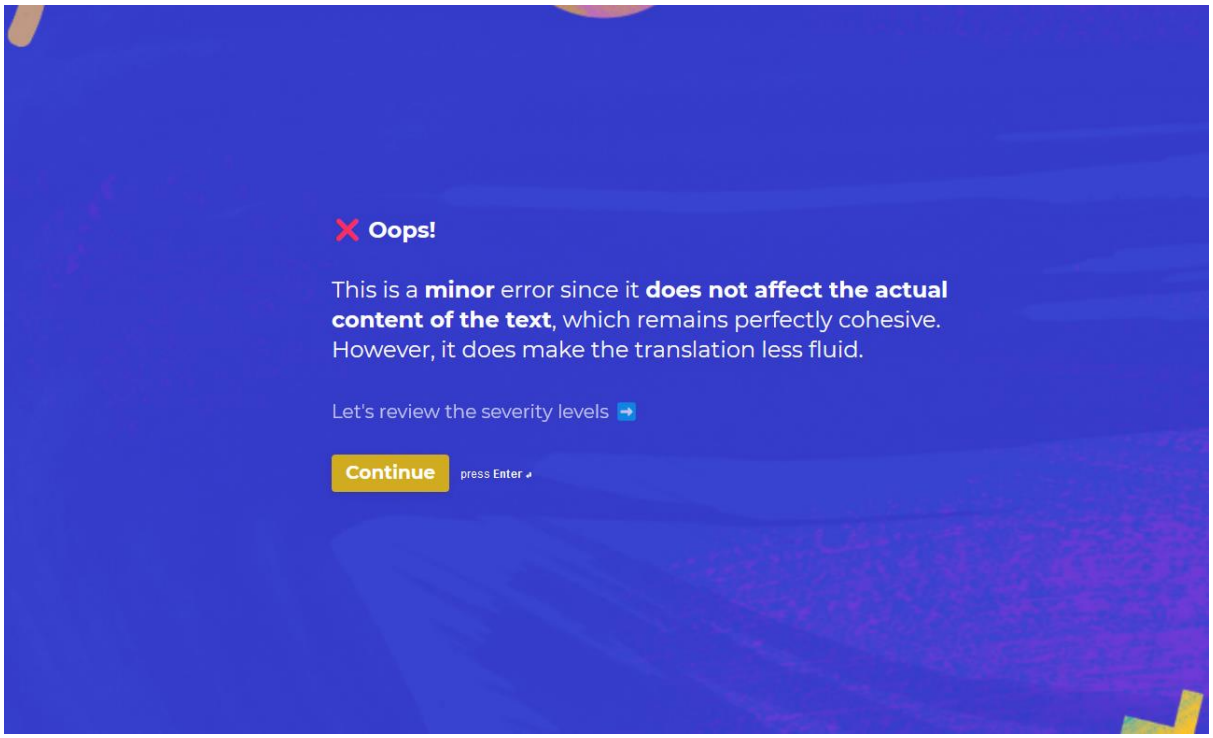

❌ Not quite.
This is a major error because it may lead the reader to not know which product the text is referring to.
While this error impacts usability or understandability of the content, it is not impossible for the reader to understand the message conveyed by the text.
Let's review the severity levels ➡️
Continue
press Enter

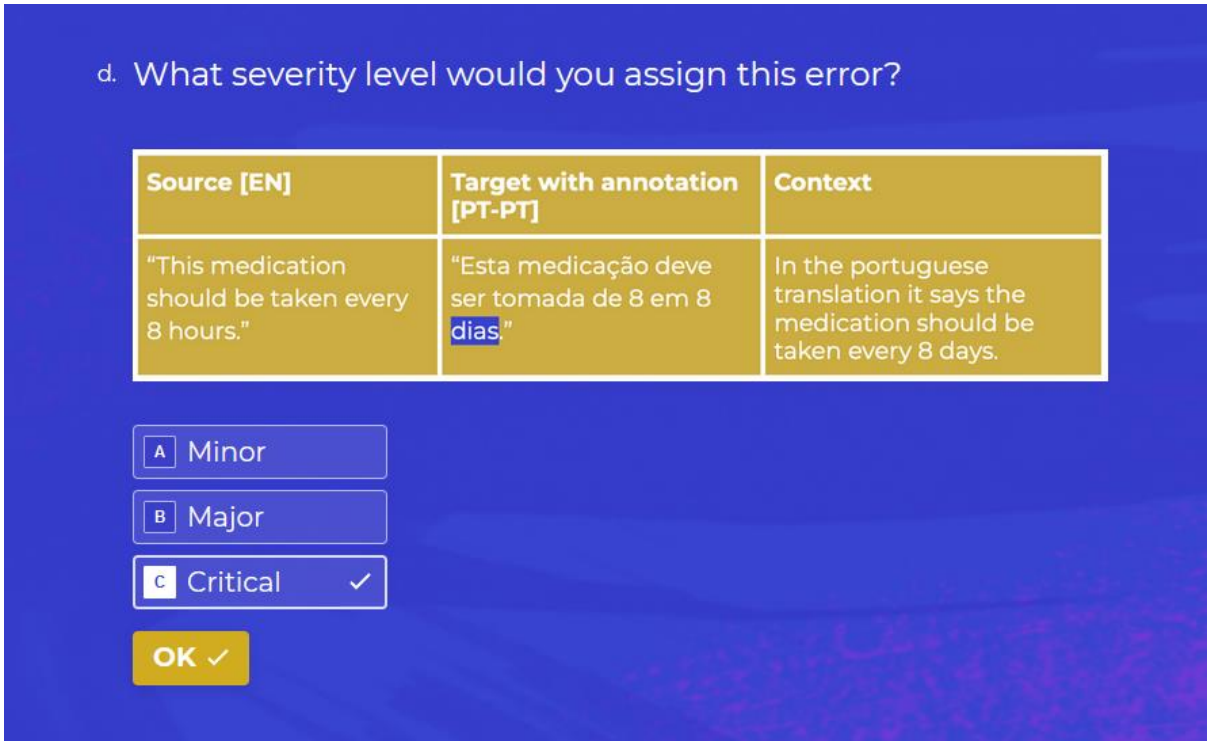

🟡 Minor
If an error does not cause the user to become confused or misled and does not result in a loss of meaning, it should be classified as minor.
Minor errors may, however, reduce the text's stylistic quality, fluidity, or appeal.
If an error does not impact the actual content of the text and only makes it less fluid, it is considered a minor error.
⚠️ Minor errors are highlighted in yellow in the Annotation Tool.
Continue
press Enter

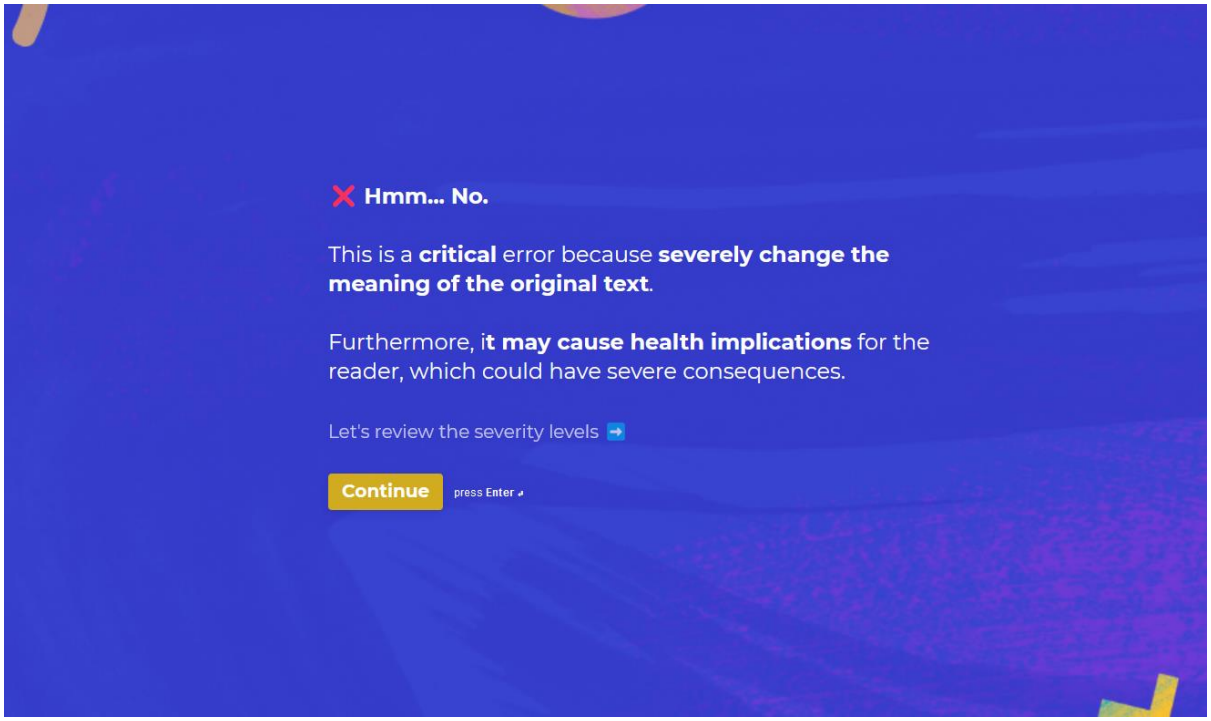

Major
Major issues impact usability or understandability of the content but do not render it unusable.
They may be difficult but not impossible to understand. This type of errors appear in a visible or important part of the text which may cause some meaning to be lost. However, it is not impossible for the reader to understand the message conveyed by the text.
Errors due to non-compliance with Company style requirements render it automatically unfit for purpose, so they should always be always assigned at least a Major severity.
Major errors are highlighted in orange in the Annotation Tool.
Continue
press Enter

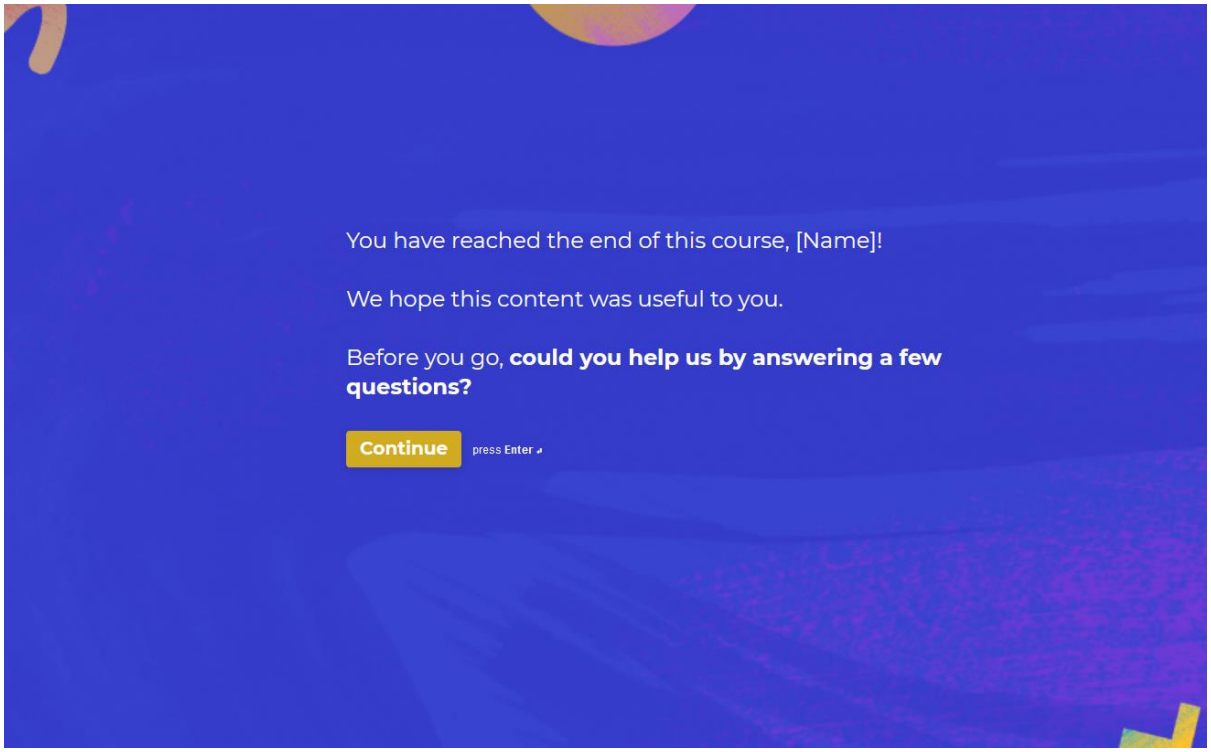

Critical
Critical issues severely change the meaning of the original text, making it impossible for the reader to understand its actual meaning.
These errors may carry health, safety, legal or financial implications to the end user/reader (e.g. if a medical information is mistranslated). They may also badly reflect on the client's reputation or violate geopolitical usage guidelines, as well as make the text come across as offensive towards an individual or a group (a religion, race, gender, etc.)
Critical errors are highlighted in red in the Annotation Tool.
Continue
press Enter

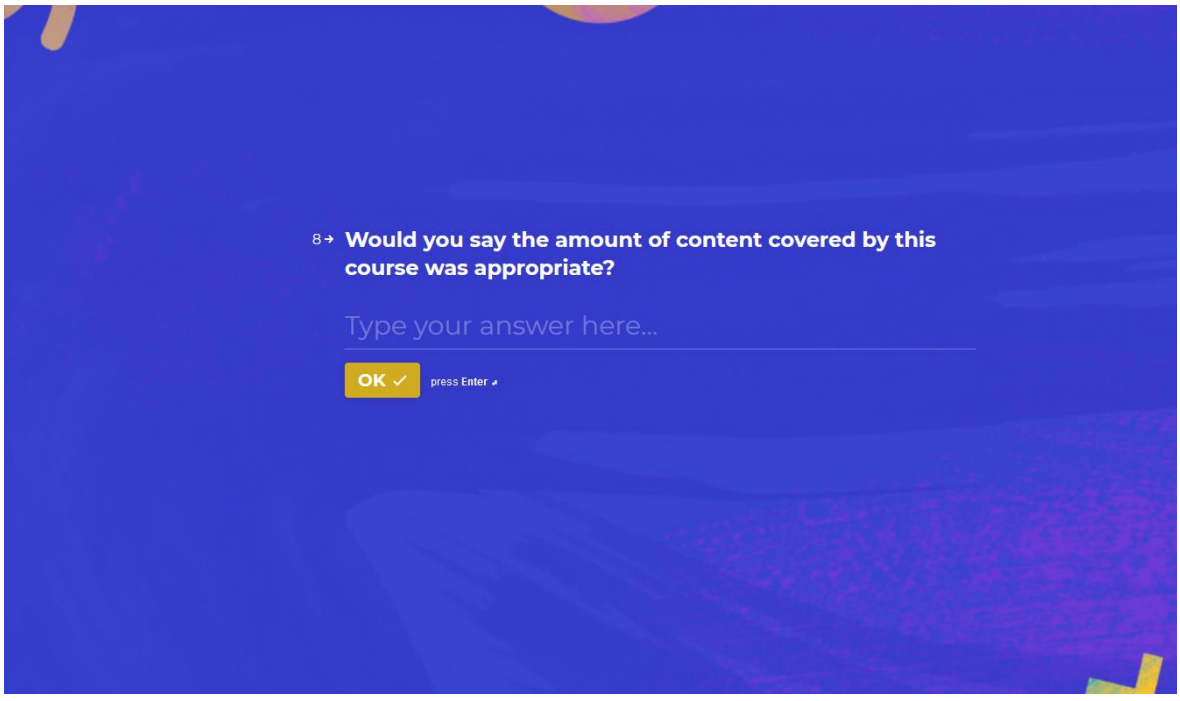

c. What severity level would you assign this error?
Source [EN]
Target with annotation [ES-ES]
Context
"TE Connectivity's (TE) Heavy Duty Sealed Series Mini 2-Position (2Pos) Connectors"
"Los conectores mini de 2 posiciones (2Pos) de la serie sellada de alta resistencia de TE Connectivity"
The word order in "conectores mini" is not very fluid, "mini conectores" would be more appropriate.
A Minor
B Major
C Critical
OK

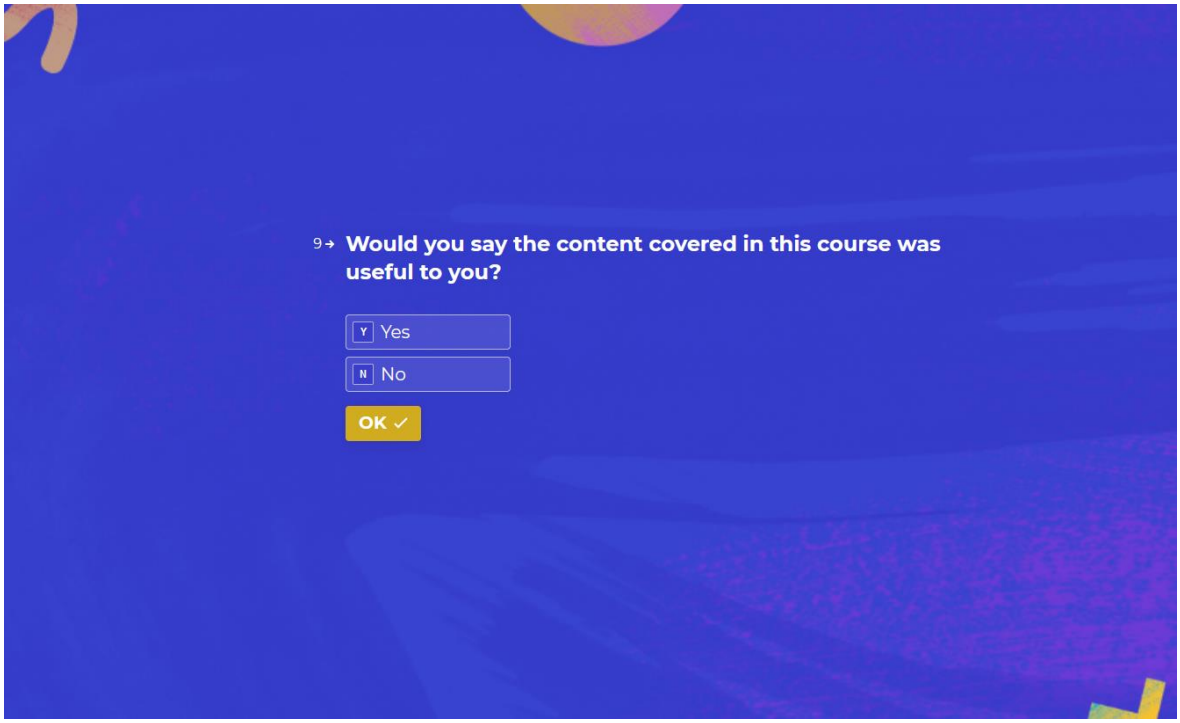

Oops!
This is a minor error since it does not affect the actual content of the text, which remains perfectly cohesive. However, it does make the translation less fluid.
Let's review the severity levels
Continue
press Enter

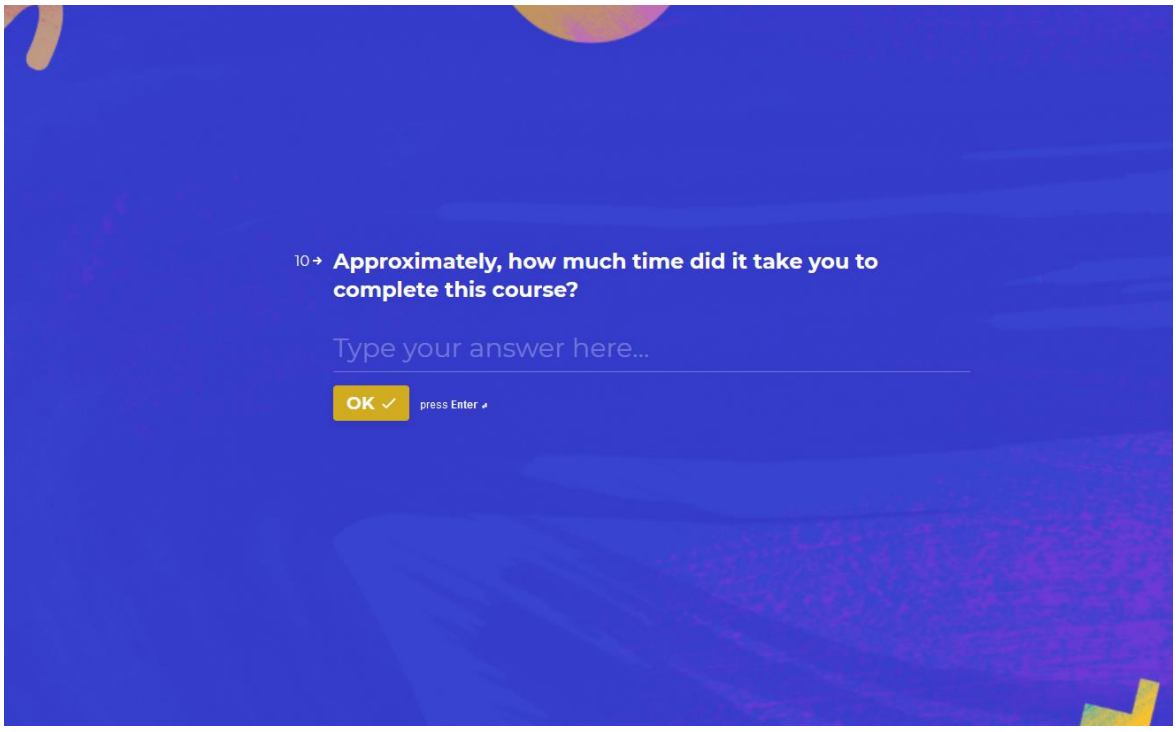

d. What severity level would you assign this error?
Source [EN]
Target with annotation [PT-PT]
Context
"This medication should be taken every 8 hours."
"Esta medicação deve ser tomada de 8 em 8 dias."
In the portuguese translation it says the medication should be taken every 8 days.
A Minor
B Major
C Critical
OK

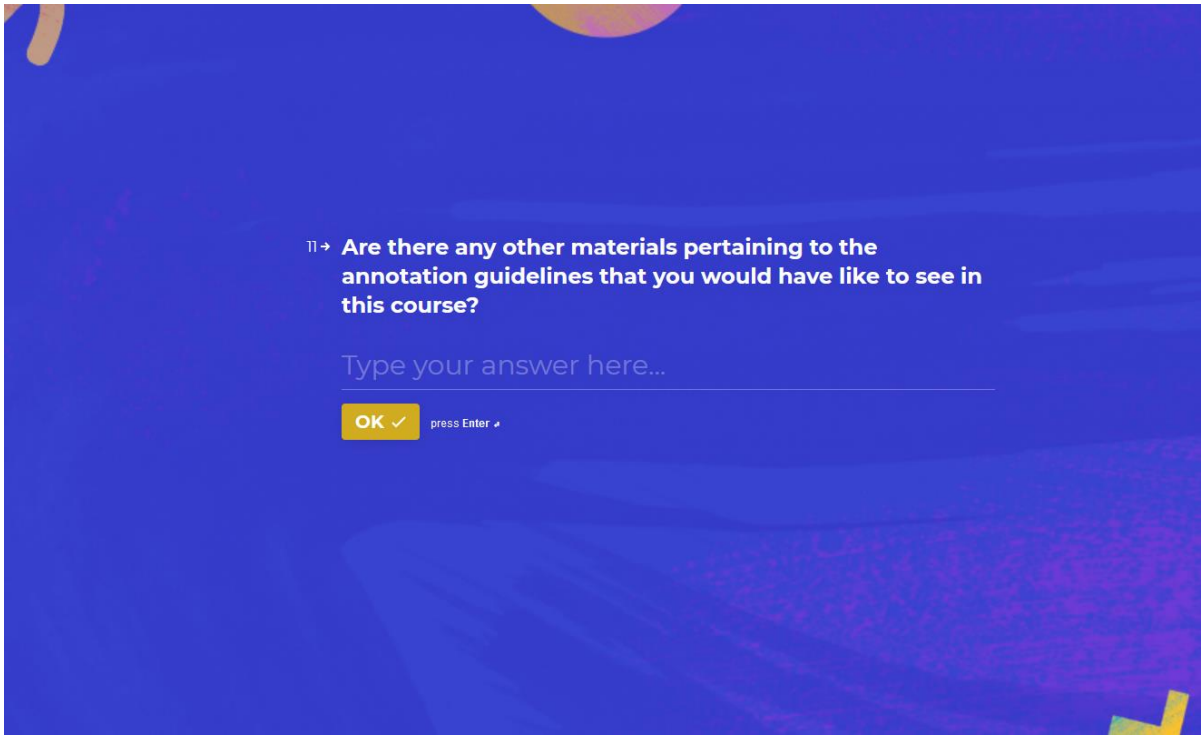

Hmm... No.
This is a critical error because severely change the meaning of the original text.
Furthermore, it may cause health implications for the reader, which could have severe consequences.
Let's review the severity levels
Continue
press Enter

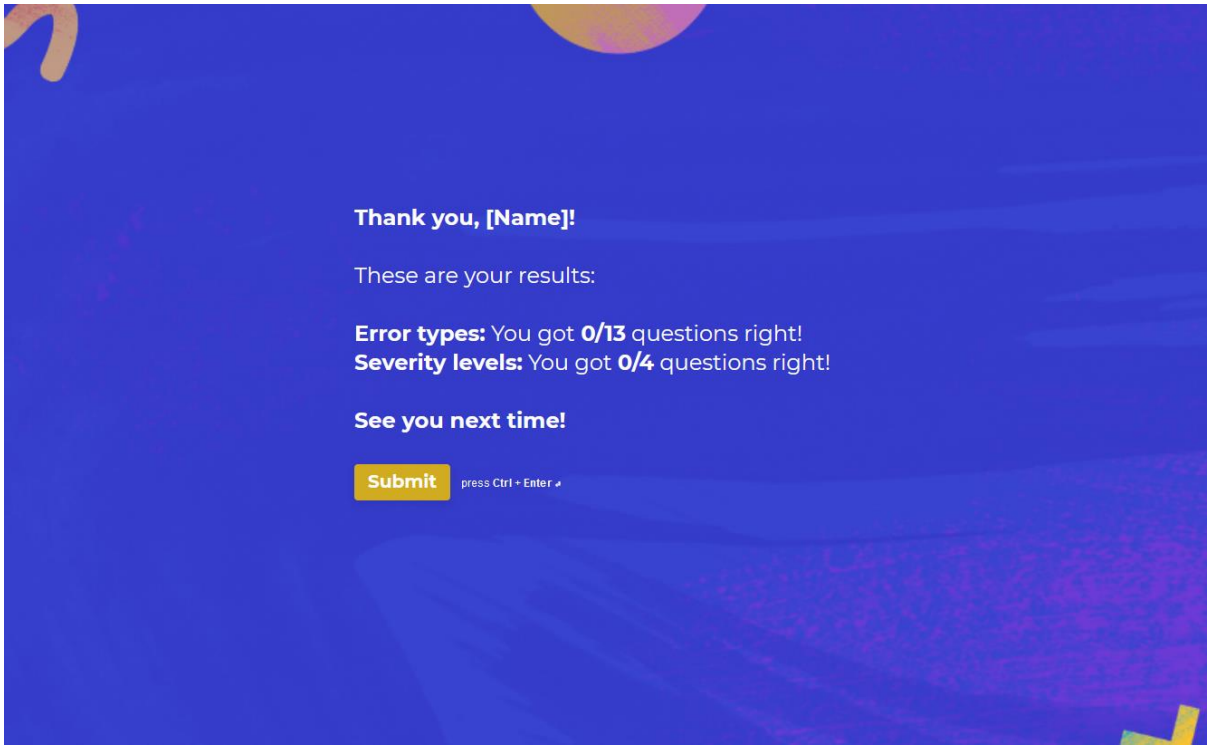

You have reached the end of this course, [Name]!
We hope this content was useful to you.
Before you go, could you help us by answering a few questions?
Continue
press Enter


8→ Would you say the amount of content covered by this course was appropriate?
Type your answer here...
OK ✓
press Enter


9→ Would you say the content covered in this course was useful to you?
Y Yes
N No
OK ✓


10→ Approximately, how much time did it take you to complete this course?
Type your answer here...
OK ✓
press Enter


11→ Are there any other materials pertaining to the annotation guidelines that you would have like to see in this course?
Type your answer here...
OK ✓
press Enter


Thank you, [Name]!
These are your results:
Error types: You got 0/13 questions right!
Severity levels: You got 0/4 questions right!
See you next time!
Submit
press Ctrl + Enter

Error types explained in the training

Mistranslation

This type of errors occur when the target text contains incorrect translation choices, such as a word, multiword expression[25] or a larger portion of text:

- doesn't fully reflect the meaning of the source.
- sounds odd or unnatural in the target language, even if the meaning of the source text is conveyed.

Wrong Named Entity

This tag should be used on any error that concerns a named entity (including mistranslated, untranslated, unnecessarily translated, wrongly transliterated, omitted or added named entities).

What is considered a named entity at Unbabel:

- People's names (including surnames, aliases and usernames)
- Company, team and product names (including movies, songs, TV shows, books and other publications, art pieces…)
- Country, city and all sorts of location names
- Email addresses and URLs
- Numerical and alphanumerical entities (including currency and measurements, phone numbers, credit card numbers, passwords, reference codes…)
- Date and time expressions
- Postal addresses

Punctuation

Punctuation errors occur when a punctuation mark is used incorrectly or is missing from the translation.

[25] A multiword expression is an expression made up of two or more words that is perceived as a semantic unit. (e.g. *credit card*, *make up*, etc.)

Addition or Omission tags should not be used to label added or omitted punctuation marks[26].

Whitespace

This type of errors occur when:

- One or more whitespaces are incorrectly added or omitted
- Two words that should be written as a single word (with or without a hyphen) are separated by a whitespace[27]
- Two words that should be separated by a whitespace are written as a single word (with or without a hyphen)

It is important not to use the Whitespace tag when the error falls under Locale Conventions. Whitespace should not be used if there is an extra (or omitted) whitespace in one of the following:

- Address
- Currency
- Date/time
- Measurement
- Number
- Telephone number

Markup Tag

Markup Tag errors occur when there are incorrect markup tags or tag components in the target text.

These errors may include malformed HTML characters and emojis where the emoji format is different from the source text's.

Well-formed escaped HTML characters should never be tagged with the Markup tag. This tag should only be used when these markups have an error in them.

[26] This also applies to the Whitespace tag.
[27] Both of these cases are only Whitespace errors if none of the words are a named entity or a glossary entry.

Term Not Applied

This category is used when the term present in the target text is not compliant with the one specified in the glossary. Term Not Applied should be used when there is a glossary term that should be applied in a certain situation but instead, another one is used.

Wrong Term

This category should be applied when the term present in the target text is compliant with the glossary, but it doesn't fit in context or there's an error in it (it can be a typo, a capitalization issue or any grammatical error).

Do Not Translate

This type of errors occur when a unit[28] was translated but it should have been left untranslated according to the client's preferences.

This tag should only be used when there's a specific client instruction regarding the need to not translate certain words, phrases or expressions.

Lacks Creativity

The Lacks Creativity label must be used when the translated text is correct and a close and true reflection of the source content, but it lacks language creativity and flexibility and is not linguistically appealing or engaging enough.

This category can only be used when creativity is an express requirement of the customer.

Unnatural Flow

[28] a word, a multiword expression or a phrase

This type of errors cover situations where a portion of text, larger than a single word or multiword expression, is a too literal translation of the source. The meaning of the source comes through in the target, but the overall feeling of the translation is unnatural.

Contrary to Mistranslation, Unnatural Flow is a matter of style that does not affect meaning.

Locale Convention

This type of errors violate locale-specific content or formatting requirements. There are six error types that fall under this category:

- Address Format (345 California St Suite 600 & 700 vs. 345 California Suite 600 & 700 St)
- Currency Format (e.g. $10 vs. 10 $)
- Date/Time Format (e.g. dd/mm/yyyy vs. mm/dd/yyyy)
- Measurement Format (e.g. 50mm vs. 50 mm)
- Number Format (e.g. 10% vs 10 %)
- Telephone Format (e.g. (211) 555-2781 vs. 211 555 2781)

Culture-specific Reference

This type of errors cover cases where the target text contains a culture-specific reference that’s not appropriate or understandable to the intended target audience.

An example of this would be the use of jargon related to sports or other culture-specific features that are not necessarily understood in the environment of the target language.

Wrong Language Variety

This type of errors occur when the language variety used is not the one requested.

An example of this would be using Brazilian Portuguese spelling or word choices in a European Portuguese text.

Source Issue

This tag must be used always in combination with another error tag, when there is an error in the target text that is due to an issue in the source text.

In this case, two separate annotations should be made: one as the actual error and one as Source Issue.

The Source Issue tag should always be assigned a Neutral severity.