Context is everything: a context-aware annotation typology for dialogue translation quality assessment

Miguel Menezes
Amin Farajian
Helena Moniz
João Graça

Miguel Menezes
University of Lisbon; FLUL/CLUL, INESC-ID, TransPerfect;
lmenezes@edu.ulisboa.pt;
ORCID: 0000-0003-4253-2534

Amin Farajian
TransPerfect;
amin@transperfect.com;
ORCID: 0000-0001-6384-5332

Helena Moniz
University of Lisbon; FLUL/CLUL;
helena.moniz@edu.ulisboa.pt;
ORCID: 0000-0003-0900-6938

João Graça
Instituto Superior Técnico, heyL;
gracaninja@heyl.ai;
ORCID: 0000-0001-7889-5332

Abstract

Until recently, most machine translation (MT) systems translated sentences in isolation, neglecting crucial document-level context due to limited discourse-focused training data and a lack of robust evaluation methods. We introduce a context-aware annotation framework, validated on a customer support dataset with substantial inter-annotator agreement (Cohen's κ = 0.73), potentially offering a new standard for contextual MT assessment.

Keywords: machine translation; discourse phenomena; context; translation quality workflows; context-aware annotation framework.

Resumen

Hasta hace poco, la mayoría de los sistemas de traducción automática (TA) traducían las oraciones de forma aislada, pasando por alto un contexto crucial a nivel de documento debido a la escasez de datos de entrenamiento centrados en el discurso y a la falta de métodos de evaluación sólidos. Presentamos un marco de anotación sensible al contexto, validado sobre un conjunto de datos de atención al cliente con un acuerdo interanotador sustancial (κ de Cohen = 0,73), que podría ofrecer un nuevo estándar para la evaluación contextual de la TA.

Palabras clave: traducción automática; fenómenos discursivos; contexto; flujos de trabajo de evaluación de la calidad de la traducción; marco de anotación. sensible al contexto

Resum

Fins fa poc, la majoria dels sistemes de traducció automàtica (TA) traduïen les oracions de manera aïllada, i deixaven de banda un context clau a nivell de document a causa de l'escassetat de dades d'entrenament centrades en el discurs i de la manca de mètodes d' avaluació sòlids. Presentem un marc d ' anotació sensible al context, validat sobre un conjunt de dades d ' atenció al client amb un acord interanotador substancial (κ de Cohen = 0,73), que podria oferir un nou estàndard per a l ' avaluació contextual de la TA.

Rebuda: 30 de abril de 2025| Acceptació: 14 de novembre de 2025 | Publicació: 15 de desembre de 2025

Paraules clau: traducció automàtica; fenòmens discursius; context; fluxos de treball d'avaluació de la qualitat de la traducció; marc d'anotació sensible al context.

1. Introduction

Recent technological advancements have driven global transformations, increasing the demand for translation into native languages. This has highlighted the need for more translators while reshaping the translation landscape, as the internet provides unprecedented freedom for translators. As a result, language service providers (LSPs) and researchers have adopted data-driven technologies such as machine translation (MT) systems that, despite showing competitive quality in standard benchmarks, are still "perceived as much worse when evaluated on entire documents rather than at the sentence level" (Petrick et al., 2023: 375). A persistent lack of document-level training corpora and the absence of clear context-aware quality assessment (QA) schemes continue to delay progress. Against this backdrop, we present a validated annotation framework, applied to a bilingual customer support dataset, that focuses strictly on document-level phenomena, i.e. context, and is compatible with Multidimensional Quality Metrics (MQM) (Lommel et al., 2014). The framework unifies (i) contextual source triggers, (ii) a contextual MT error typology, (iii) severity levels, and (iv) decision rules with descriptive guidelines, yielding a more comprehensive operational scheme for diagnosing context-dependent MT errors and their source sentence triggering elements. The paper proceeds as follows: Section 2 reviews related work; Section 3 presents the framework; Section 4 details data and methods; Section 5 reports on results and discusses implications and limitations; and Section 6 sets out a conclusion and outlines future work.

In the last few decades, translators have relied on extensive digital resources, such as translation memories (TMs), bilingual corpora, electronic databases, and dictionaries, to enhance quality, productivity, and consistency. More recently, AI-driven computer-assisted translation (CAT) tools like Trados have integrated MT, leveraging advances in neural modelling (Läubli et al., 2020). To scale operations, many professionals have embraced automation, with some LSPs specialising in MT services. MT has become ubiquitous, redefining translation roles and shifting translators toward QA and post-editing tasks. Despite these changes, translators continue to have an essential role as quality curators and gatekeepers, ensuring high translation standards.

MT systems excel at learning and applying linguistic patterns, resulting in them achieving high-quality results and prompting claims of human parity. However, as previously stated, automatic translations are worse when assessed at the document level than at the sentence level, exposing limitations in sentence-level strategies. Until very recently, most MT models remained indifferent to a document's overall context, translating sentences in isolation (Bawden, 2018). This is particularly detrimental to discourse mechanisms, i.e. the textual devices that add, update, and restrict information across

sentences. Failing to account for these interdependent structures undermines cohesion and coherence in the target text and compromises meaning transfer.

Aware of such limitations, developers have sought to address the bottleneck by developing MT architectures with document-level attention mechanisms to encompass multiple sentences or sequences during the translation process, allowing for discourse-related phenomena at the document level. Nevertheless, most MT systems today still operate on the sentence level. "One of the main reasons for this is that most document-level approaches rely on parallel training data with document-level metadata. Most releases of large parallel training corpora lack this information and remain purely sentence-level" (Petrick et al., 2023: 1). This lack of resources has delayed the paradigm shift in MT. Moreover, and as highlighted by Post & Junczys-Dowmunt (2023), there is also a lack of comprehensive, flexible evaluation strategies that focus on contextual phenomena. For the most part, the metrics and frameworks used to measure contextuality remain sentence-level, failing to account for contextual mechanisms and to provide an accurate portrayal of improvements in document-level MT systems (Jwalapuram et al., 2021). Given this scenario, developers strongly question the value and effectiveness of investing in these models (Yin et al., 2021; Jin et al., 2023). This makes it especially difficult to evaluate any type of model, posing a serious challenge to MT developers (Post & Junczys-Dowmunt, 2023).

Despite this bleak outlook, the current gap in context-aware MT evaluation has sparked curiosity and interest among certain members of the MT community interested in understanding the genuine capability of an MT model to handle complex contextual phenomena. This has resulted in the emergence of various MT evaluation proposals, such as annotated test sets, test suites or tool kits, which use context as a key quality metric. For a detailed discussion, see Section 3.1.

Nevertheless, despite several attempts to overcome the aforementioned challenges, there is still no one-size-fits-all QA method for context-based MT evaluation. All strategies have pros and cons, and despite great strides made in this area, it seems much work needs to be done before a consensus is reached. For now, most context-aware evaluation frameworks and metrics proposed have been shown to be reductive, with limited contextual representation, focusing mainly on a rigid set of contextual categories and paying particular attention to a single side of this equation, contextual MT errors, rather than also pinpointing their triggering elements within the source. We draw attention to this problem and highlight the need for consolidation to achieve a more realistic and fine-grained understanding of the current MT landscape.

In response to current needs, and through a data-oriented approach, our initial engagement with ecological data bridged theoretical linguistic concepts (Section 2) and empirical examples, facilitating the construction of the proposed annotation framework. For implementation of our context-aware annotation framework, we selected customer support chat MT, an idiosyncratic domain that challenges boundaries between written and spoken language. To this end, we considered a dataset in which the customer and agent communicate in their native languages, with MT facilitating seamless bidirectional

interaction. It is important to highlight that these conversations are often composed on the fly, rendering them poorly structured, occasionally ungrammatical, and potentially incoherent. Additionally, high emotional load further complicates challenges in interpretation. Context becomes essential in bridging these gaps, reinforcing its importance for accurate translation. Given these factors, this domain provides an ideal scenario to: (i) identify domain-specific discourse phenomena, (ii) apply our framework and test its ability to detect and categorise contextual MT errors alongside their source-level triggers within the source document, (iii) measure the impact of disregarding contextual information on translation quality, and (iv) evaluate MT effectiveness in real-world conditions.

2. Understanding context: theoretical insights and approaches

Communication encompasses various definitions, from information exchange to social interaction, all rooted in the idea that understanding arises from shared common ground. Introduced by Grice (1991) and developed by Stalnaker (2002), this refers to the "background information" shared by conversation participants, implicitly conveyed through the "context set", a set of possible worlds (Stalnaker, 2002: 701). Language plays a central role in this process, serving as "the specific medium of understanding at the social stage of evolution" (Habermas, 1979: 1). Crucially, communication is not built on isolated linguistic units but on their combinations, as people use structured expressions to convey meaning (Shen, 2012: 2663).

Building on this idea, communication can be seen as bidimensional, involving both explicit information, i.e. clearly conveyed content, and implicit information, where meaning is inferred beyond the words. In any sentence or text, some information is directly stated, while other aspects rely on interpretation by the listener. As Horn & Ward (2004) highlight, language provides a framework for expression, but linguistic information alone may not fully capture a message's intent. Thus, communication is more than merely "encoding and decoding messages" (Horton, 2012: 375); speakers and listeners depend on both linguistic and extralinguistic context to convey meaning effectively.

By following Horn & Ward's (2004) perspectives, such a process can be systematised in the following manner:

> The speaker's utterance (p), carrying the implication (q), is provided within a context (C); the inference of (q) is prompted by the speaker's utterance of (p) in (C).
>
> I (q | p, C) = f (Speaker (p), Implication (q, p), Context (C))
>
> I (q | p, C): This represents the inference q (the implication) given the utterance of proposition p within context C.
>
> f: This is the function that combines the following information:
>
> (i) Speaker (p): This considers details about the speaker making the utterance p.

(ii) Implication (q, p): This considers the relationship between the utterance p and the inference q. It considers semantic meaning, linguistic rules, pragmatic principles, and shared knowledge (common ground).

(iii) Context (C): This represents the context in which the utterance occurs, the background information, or the framing.

Theoretical work that deals with inference and presupposition emphasises context as crucial in shaping meaning, as reflected in the idea that "context determines text and text reflects context" (Shen, 2012: 2663). This circular dependency appears even at a micro-level, where language units take on different meanings depending on context such as homonyms and polysemous words. Wittgenstein & Anscombe (1958) argue that a word's meaning is defined by its use in language, suggesting broader contexts lead to greater semantic flexibility. As new contexts emerge, so do new meanings.

However, context extends beyond linguistic expressions, functioning as a multidimensional concept that includes: (i) cultural context, shaped by native traditions; (ii) situational context, tied to the circumstances of an utterance; and (iii) linguistic context, determined by grammar and syntax (Malinowski, 2000: 301-305). This threefold perspective underscores the interplay of language, culture, and social cues in shaping meaning and communicative success.

Building on Malinowski's (2000) three-dimensional view of context, developed by Firth (2020), we adopt this perspective to identify potential translation errors arising from unclear or overlooked contextual dimensions. To enhance this framework, we integrate the concept of dynamic semantics, which acknowledges that context is not static but evolves within a text. This theoretical model explains how utterances are "inherently dynamic and continually updat[ed] as the discourse progresses" (Birner, 2012: 227). Each new piece of information reshapes the conversational context, renewing the shared knowledge between participants. This process can be formulated as follows:

$[C + P] \rightarrow c'$

In which:

C: the common ground, i.e. shared body of information.

P: new information

c′: common ground + new information.

Thus, c′ represents the updated context and the new body of information.

By merging the two models, i.e. the three-dimensional contextual perspective along with the dynamic semantics theory perspective, in which context is continuously updated, we can account for the entire contextual spectrum in a document and measure it. Therefore, our perspective of context entails: (i) the textual components, (ii) the situation

of enunciation, and (iii) world-knowledge, in a (iv) dynamic context. Together, these components form a complete definition of what context represents, and they hold the basis of our research endeavours.

2.1. Contextual mechanisms

Previously mentioned perspectives of context in a document are conveyed through contextual mechanisms, or contextual proxies, which support textuality, i.e. how a text is structured and interpreted (Silverman, 1986). These proxies uphold coherence and cohesion, the fundamental parameters that shape meaning by linking parts of a text (Halliday & Hasan, 1989).

While interrelated, coherence and cohesion have distinct roles (Bublitz, 2011). Cohesion refers to explicit linguistic links between sentences, detectable at the surface level. Coherence, in contrast, is conceptual, relying on sociocultural context, communicative intent, and shared knowledge. A text is coherent when it is comprehensible and acceptable.

In discourse analysis, cohesive mechanisms are central to textuality, extensively explored by Halliday & Hasan (1989) and Tierney & Mosenthal (1983). Referentiality, a key mechanism, involves anaphoric, deictic/situational, and epistemic dimensions. Its inadequacies can lead to ambiguity and hinder comprehension. For illustration, consider the following example, which underscores the critical role of referentiality in discourse to access meaning.

Example 1

The President has just been convicted. The trial results have restored a renewed sense of trust in the country's judicial system.

Identifying the referents of "The President" and "the country" requires prior discourse or situational context, neither of which is explicitly given. However, using extralinguistic knowledge, one can infer that the country is a democracy with a reliable judicial system. Ellipsis is another key mechanism, where part of a sentence is omitted but remains understandable through context. Halliday & Hasan (1989: 142) describe this as "substitution by zero" and explain that ellipsis involves presupposition — an assumption about what has been left out. Consider the following example of ellipsis provided by the authors:

Example 2

"And how many hours a day did you do lessons?" said Alice, in a hurry to change the subject.

"Ten hours the first day [-]", said the Mock Turtle: "nine [-] the next [-], and so on."

The omitted elements (e.g. "hours") are understood from the previous sentence, showing how meaning relies on context. The same applies to the nominal group beginning with "next", as in "the next day". These mechanisms illustrate how context shapes communication. In MT, they provide a framework for assessing contextuality, ensuring translations maintain coherence across a document. Next, we explore translation quality within the scope of Translation Studies (TS), with a particular focus on recent MT QA methods and frameworks from a contextual perspective.

3. Translation quality assessment

Before evaluating translation quality, we must first: (i) define quality, (ii) determine its relevance to our project, and (iii) identify available assessment tools. Koby et al. (2014) emphasise the need for an explicit definition of translation quality to ensure an objective, practical, and realistic approach to error identification and correction. In *Defining Translation Quality*, the authors acknowledge the complexity and ambiguity of the concept, arguing that no universal definition exists. Instead, they propose two contrasting perspectives: a broad and a narrow definition of translation quality. The former perspective states:

> A quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs. (Koby et al., 2014: 416)

This perspective on quality follows a functionalist approach, emphasising the translation's ability to fulfil its intended purpose for its target audience. Achieving this goal requires communication between the translation requester and provider. Translation, seen as an ecological system (Fang, 2018), is shaped by external factors such as economy, language, culture, and politics. Vermeer's *Skopostheorie* (1978), grounded in Action Theory (*Handlungstheorie*, Von Wright, 1981), frames translation as a purposeful action (Nord, 2014: 12), driven by intent and motivation.

Integrating Koby et al.'s (2014) view on quality with Vermeer's action theory enhances effectiveness by: (i) ensuring balanced roles among translation stakeholders, (ii) fostering a more integrative and structured translation process, and (iii) improving overall translation quality. This leads to a well-structured process accounting for pre-translation, translation, and post-translation stages, maximising the likelihood of high-quality results.

Returning to the broader definition of quality, the concept presented by Koby et al. (2014) aligns with the views of Garvin (1984). Garvin, in his article *"What Does Product Quality Really Mean?"*, introduces a five-part approach to quality definition dependent on different angles that encompass philosophical, product-based, user-based, marketing- and management-based, and value-based perspectives. Garvin's view of quality as "innate excellence" and "a mark of uncompromising standards and high achievement" (Garvin, 1984: 25) was transferred to TS through a set of formulas that allow objective measurement of quality in translation, ensuring compliance with the author's standards.

According to Koby et al. (2014), Garvin's product-based approach, in which "differences in quality reflect differences in the quantity of some ingredient or attribute possessed by a product" (Garvin, 1984: 25-26), was applied to translation quality assessment (TQA) using a document's accuracy and fluency values as measurable criteria to gauge quality. We have based the proposed annotation framework on this approach in terms of analysis of a document's contextual variables (ingredients), expressed through quantifiable "contextual" categories. We will explore this matter in more detail in the following sections.

As stated, the authors also found it necessary to present a second, contrastive perspective, consisting of a narrower view of quality:

> A high-quality translation is one in which the message embodied in the source text is transferred completely into the target text, including denotation, connotation, nuance, and style, and the target text is written in the target language using correct grammar and word order, to produce a culturally appropriate text that, in most cases, reads as if originally written by a native speaker of the target language for readers in the target culture. (Koby et al., 2014: 416-417)

In this definition of quality applied to translation, the focus lies on the text. It is thus "text-centric", with minimum absolute requirements, which can be thought of as established baselines that are non-negotiable for the work to be considered of a high standard.

These two perspectives on translation quality do not cancel each other; on the contrary, they are complementary. While one considers defining operational strategies and translation quality management, often handled with the assistance of language operations (LangOps) teams, the other, with its narrow scope, refines the concept of translation quality to the linguistic reality present in the source document and its reflection in the target document. This perspective shifts attention from broader considerations to a more detailed, linguistically oriented examination of the relayed information within both documents.

Both quality perspectives are essential to our approach to MT QA. Our context-aware annotation framework evaluates co-textual elements, such as anaphoric and cataphoric relationships, aligning with the narrow perspective. It also considers extralinguistic factors, like user expectations in customer support, reflecting the broader definition. However, our primary focus remains the linguistic quality of the bitext.

In the following sections, we provide an overview of MT QA approaches, ranging from manual evaluation to advanced automatic techniques. These methods have strengthened and supported MT quality over time, since "[i]mplementing quality assessment methods is essential to monitor the evolution of MT systems" (Escribe, 2019: 36). We also show how our proposal aligns with current MT trends, facilitating dataset annotation that can serve as a leverage point to enhance or refine existing automatic MT QA metrics, enabling more effective context-based assessments.

3.1. Context awareness in MT assessment: gaps & challenges

TQA is inherently complex and subjective, further complicated by the elusive definition of translation, i.e. a cognitive, linguistic, social, cultural, and technological process (Castilho et al., 2018). Different theoretical perspectives shape translation approaches, ranging from word-for-word (*verbum pro verbo*) to meaning-based (*sensum de sensu*) methods, with some emphasising audience-oriented strategies (Nord, 2014). A text-centric approach, as proposed by Koby et al. (2014), offers a solution by enabling empirical evaluation of MT quality through human annotation or automatic metrics focused on meaning transfer, grammatical correctness, and linguistic accuracy.

Context plays a crucial role in communication, but has long been neglected in MT due to past technological limitations, by the misconception that sentence-level neural machine translation (NMT) had reached human-like quality (Hassan et al., 2018), and the scarcity of document-level training and evaluation resources, all of which are still a major barrier to context-aware MT (Wicks & Post, 2023).

Empirical studies have debunked claims of human parity in MT (Läubli et al., 2018; Toral et al., 2018), underscoring the need for refined evaluation protocols. While context-aware MT is gaining traction, challenges persist, including limited document-level data, preprocessing methods that strip context, increased latency compared to sentence-based MT, and the lack of comprehensive evaluation frameworks (Post & Junczys-Dowmunt, 2023). Widely used metrics like BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) remain sentence-level, limiting their effectiveness in assessing document-level MT improvements (Jwalapuram et al., 2021).

Despite these obstacles, interest in context-aware MT evaluation is growing, driving the development of new methods for assessing MT's handling of contextual phenomena, though much work remains.

3.2. Discourse-based MT evaluation methods

As previously stated, the last decade has seen the production of a significant body of work aimed at developing a trustworthy method for assessing MT quality from a contextual standpoint, with the appearance of several MT benchmark test sets and test suites for context-based MT evaluation, such as the large-scale test set for the evaluation of context-aware pronoun translation in NMT proposed by Müller et al. (2018) or the test suite for evaluating discourse phenomena in document-level NMT proposed by Cai & Xiong (2020). These annotated datasets play a crucial role for developers, as they provide a ready-made and easy-to-use method for MT evaluation which can be elicited at various stages of development, despite their limited MT error coverage. Following a similar evaluation design, Castilho et al. (2021) developed a document-level corpus annotated, the DELA Corpus, a document-level corpus annotated with context-related issues for En→Pt/Br.

Besides the above solutions, other methods have been developed for evaluating contextual MT, particularly automated metrics and accessible toolkits, an example being

the work of Vernikos et al. (2022). The authors proposed an alternative method for incorporating context into four pre-trained MT evaluation metrics: BERTscore, which uses large language models (LLMs) to score generated text, Prism, a text generation metric, COMET, an embedding-based metric and the reference-free COMET-QE. Their aim was to improve these metrics' ability to handle ambiguities by adding contextual information. To evaluate how well the extended metrics aligned with human MT quality assessments, they used WMT 2021 MQM annotations (Lommel et al., 2014), on the grounds that:

> MQM judgements are the best available to test document-level MT metrics, as these judgements are made by expert translators that have access to and are strongly advised to consider source-side document-level context when judging each target sentence. (Vernikos et al., 2022: 118)

MQM scores are often used to measure correlation under the assumption that:

> Results are produced by professional translators (compared to crowd workers or translation researchers) and require explicit error annotations that are believed to lead to higher quality annotations [and that] MQM annotators are specifically instructed to identify all errors within each segment in a document, paying particular attention to document context. (Vernikos et al., 2022: 122)

However, as Menezes et al. (2023) noted, MQM results should be interpreted with caution in document-level annotation. In their analysis of two MQM-annotated En→Br/Pt MT outputs for the WMT 2022 shared task on chat translation, they observed:

> Core MQM typology used for the WMT-2022 chat shared task moderately identifies some contextual issues, in part because annotators were instructed to, if possible, account for some dependencies within the dataset. Nevertheless, 36.1% for the Baseline and 42% for Unbabel-IST of the contextual issues annotated by the Context-Aware Typology were not considered during the WMT-2022 chat shared task MQM annotation. (Menezes et al., 2023: 293)

The MQM framework typology is a well-structured, widely used hierarchical error classification system that can be easily adapted by users. It includes 100 issue types with varying granularity. However, it was not originally designed to capture contextual MT errors. Applying it in expectation of it fully reflecting textual contextuality reveals its limitations. Still, it has been applied in the evaluation of context-aware MT models (Freitag et al., 2021; 2022), potentially yielding unreliable results. This gap motivated us to design an MQM extension that accounts for contextuality, allowing for more thorough evaluation of both MT and human outputs. As mentioned, we also used the same dataset to compare MQM annotations with results from our context-aware framework in its initial phase, showing the benefits of using a dedicated framework for context. Results appear in Menezes et al. (2023).

3.3. Context-aware annotation

Considering the limitations and inadequacy of current MT evaluation methods, we propose a context-focused framework over mainstream QA metrics for document-level NMT assessment, potentially resolving earlier challenges. Our framework is to be used taking

both the MT error and its triggering elements in the source text into account; *de facto*, it is the interplay between MT error and source trigger element that makes it possible to recognise whether an MT error has a contextual basis or not. Before we present the full context-aware annotation framework, however, it is crucial to define key concepts that are fundamental to its proper implementation.

3.3.1. Defining a contextual MT error

Within the scope of this project, we are singularly concentrating on MT errors that arise from insufficient contextual information within the source sentence, jeopardising meaning transfer and compromising textual coherence and cohesion. In this work, we define a contextual MT error as a mistranslation caused by a source lexical structure that relies on context but lacks enough information within the source sentence for an accurate translation. To identify contextual MT errors, it is key to consider both the source and the target text. A potential contextual MT error can be identified by looking at the entire document or by examining a specific range of sentences. The following example (extracted from an ecological dataset presented in later sections) highlights this characteristic, showing a coherence translation disruption:

> Example 3
>
> Source: While your account is on pause, you will not be *billed* for a new month of subscription.
>
> Target: Enquanto a sua conta estiver em pausa você não será *cobrad(o)/ cobrad(a)* para um novo mês de assinatura.

Example 3 shows an agreement MT error. The words in italics follow language conventions bound to a gender agreement in Pt/Br (the target sentence) but not in En (the source sentence). Since there is no information within the source sentence that allows identification of the addressee's gender, previous or subsequent contextual clues in the complete document must be followed to properly identify the head of the referential chain, i.e. the entity to which *cobrado/a* (Pt/Br) / *billed* (En) refers.

3.3.2. MT error severity

In addition to categorising MT errors, the context-aware annotation framework provides a final MT quality score, essential for comparing different MT outputs. Our approach follows Lommel et al.'s (2014) methodology, utilising a "weighted scoring" system where each MT error is assigned a severity level:

> Minor: errors that do not alter meaning but affect fluency or style.
>
> Major: errors that impact comprehension, making the translation difficult to understand.

Critical: errors that severely change meaning, have serious consequences (e.g. legal or safety risks), or cause offense.

Each MT error is tagged based on these severity levels, and a scoring formula is applied to assess overall contextual MT quality at the sentence level, following in Lommel et al.'s (2014; 2024) footsteps:

100–(sum(Minor)+((Major)∗5)+(sum(Critical)∗25))/sum(total MT Words)∗100

This formula remains effective even in a scenario where an MT hypothesis contains a smaller number of errors. In such cases, if these errors are critical, although fewer in number, they will have a greater impact on the overall MT quality score, reflecting their severity and lowering the score accordingly.

3.3.3. Defining a source trigger

As originally observed by Menezes et al. (2023), contextual triggers are lexical units within sentences which, to be processed properly, require access to information that is often situated elsewhere in a document. Examples of common contextual triggers include referential dependencies, where the same entity is referred to throughout a document using different strategies to avoid repetition, such as the use of pronouns. Look again at Example 3, a referential chain disruption due to the source lacking the necessary information for gender disambiguation. Note that the past participle is gender-marked in Pt/Br but not in En, which triggers the MT error. To accurately translate the example above, a translator would typically rely on co-textual cues.

The majority of existing contextual evaluation frameworks tend to place disproportionate emphasis on MT errors. It is important to go a step further and consider both the MT error and the source trigger as integral components of a unique process. This is particularly relevant to determining contextuality, since context, or lack of it, can be better determined within the source; the MT error is just an expression of the source sentence condition.

3.4. Context-aware annotation framework

The full proposed context-aware annotation framework features nine categories to be used on the source side, corresponding to contextual triggers; and six categories to be used on the target side, corresponding to the errors that the MT system outputs when translating the source triggers. The contextual trigger categories to be applied on the source side are as shown in Table 1.

Category	Example and Explanation
Discourse Marker: Fillers or other words that are used to indicate dialogue interactions. Different discourse markers convey different meanings for the fluidity of a dialogue.	*Source*: Thank you please try the following steps: *Target*: Obrigado, por favor, tente os seguintes passos: *Source*: Delete cache, restart your device. *Target*: Delete cache, reiniciar o seu dispositivo *Source*: ***Ta´ bom*** *Target*: ***It is good***
	Explanation: The expression **Ta´ bom** should have been translated as an acknowledgement discourse marker, such as ok, instead it is literally translated as It is good.
Ellipsis: Refers to omission of word(s) within a sentence. Syntactically, the linguistic information is recovered	*Source*: It looks like this inquiry requires further investigation, and we'll need to log into a few different systems. *Target*: Parece que esta pesquisa requer mais investigação e precisaremos de entrar em alguns sistemas diferentes. *Source*: Quando ***[-]*** forem consultar a principal questão é sobre os créditos não expirarem mais *Target*: When ***they*** go to consult, the main question is about the credits do not expire more
	Explanation: the elliptical pronoun [-], wrongly translated as they, is only recovered accessing previous sentences: "we'll need to log into a few different systems". Correct translation: When you go to consult (...).
Greetings: Conventionalised expressions used as part of our daily lives when greeting, well-wishing and leaving a conversation. These structures are dependent on the degree of politeness and cultural awareness.	*Source*: ***Bom dia***. *Target*: ***Good day***. *Source*: Gostaria de saber melhor como funciona os créditos. *Target*: I would like to know better how the credits work.
	Explanation: The expression ***Bom dia***, can be translated in En as ***Good day*** meaning it is a good day, but it should have been translated as a greeting **Good morning**, Hello. Since greetings are culturally and language dependent, they are negatively influenced when contextual information is scarce.
Lexical ambiguity: Refers to the polysemy of words in distinct contexts.	*Source*: Is there anything else I can help with? *Target*: Há mais alguma coisa em que eu possa ajudar? *Source*: ***Aside*** from that one? *Target*: ***Ao lado*** disso?
	Explanation: Aside is polysemic, meaning "at theside of x", "besides" or "apart from". The MT system translated aside as "at the side of x", instead of "apart from" due to lack of contextual information access. It needed to basis its decision on the previous sentences.
Multiword Expressions: Compounded units, for example phrasal verbs, act as a single unit. These structures can either be solved within a sentence or require contextual information to be disambiguate.	*Source*: Cancelei meu plano mas mesmo assim me cobraram. *Target*: I cancelled my plan but still they charged me. *Source*: Thank you for reaching #PRS ORG#! *Target*: Obrigado por entrar em contacto com #PRS ORG#! *Source*: Let me check that for you. *Target*: Deixe-me verificar isso para você. *Source*: Please hold while I ***pull up*** your account. *Target*: Por favor, mantenha enquanto eu ***retirei*** sua conta.
	Explanation: The Multiword-expression ***pull up*** was translated as ***retirar***, meaning: to withdraw. However, in the specific context the correct translation would be: enquanto acesso à tua conta (glosa: whilst I access your account).
Named Entity (NE): Linguistic structures which refers to, e.g., a book title, a person's name, an address, a credit card number.	*Source*: Boa tarde, não consigo comprar livros com nenhum cartão de crédito apenas com cartão de oferta. *Target*: Good afternoon, I can't buy books with no credit card only with offer card. *Source*: O último foi hoje, à pouco e chama-se ***a única mulher*** *Target*: The last was today, shortly, and it is called ***the only woman***
	Explanation: The NE title, in red, is not identified as such, and should not have been translated, since the user is looking for the book in Portuguese, but the original book's name was translated.
Reference: Targets gender and number agreements.	*Source*: Por quanto tempo vou poder ficar **afastada**? *Target*: How long will I be able to stay away? *Source*: While your account is on pause, you will not be ***billed*** for a new month subscription. *Target*: Enquanto sua conta estiver em pausa, você não será cobrad***o/a*** para um novo mês de assinatura
	Explanation: Gender agreement: masculine ***cobrado***/ feminine ***cobrada*** beyond the sentence level. In the example, only by accessing previous information (context **afastada**) we are able to understand that we need the feminine translation **cobrada**.
Register: Degrees of politeness where speakers adapt their discourse according to the audience.	*Source*: How can I help ***you*** today? *Target*: Como posso ***te*** ajudar hoje?
	Explanation: In the example, help ***you*** / ajudar-***te*** is not appropriate, since it uses a very informal second-person singular. The correct translation would be as follows: Como posso ajudá-***lo/la***?, a singular third person.
Terminology: Targets terms that constitute a set of vocabulary within a specialized field of knowledge.	*Source*: On your phone or tablet, open the #PRS_ORG# app. *Target*: No seu telefone ou tablet, abra a aplicação #PRS_ORG# . *Source*: At the top right, tap More. *Target*: Na parte superior direita, clique em Mais. *Source*: Tap ***history***. *Target*: Tap ***história***.
	Explanation: Contextually, the word "***history***" is a term and should be translated as ***histórico***. In this case, the MT does not recognizes "**history**" as a term.

Table 1: Source trigger categories with corresponding examples and explanations.

As previously mentioned, the framework also considers contextual MT errors, allowing for them to be linked with contextual triggers. Table 2 displays the proposed framework's complete set of contextual MT error categories.

Category	Example and Explanation
Agreement: MT errors are frequent and more than often because the intrasentential information needed to determine the gender or number is absent from the sentence in question.	*Source:* **Obrigada**. *Target*: Thank you. *Source*: While we are no able to provide a definite number right now, if there is a particular you are **interested** in (...). *Target*: Embora não possamos oferecer um número definitivo agora, se há um estúdio específico em que você está ***interessado***.
	Explanation: There is a gender agreement issue in the target sentence ***interessado***. The source sentence does allow one to define gender-relevant information, leading to MT error. The correct gender is signalised in the previous utterance above through the word **Obrigada**.
Collocation: These expressions composed of two or more words that are often closely associated. These combinations are context-dependent and do not have a one-to-one correspondence between languages.	*Source*: It is about a **subscription plan** pricing. *Target*: É sobre o preço do plano de assinatura! *Source*: Sim para **contratação** corporativa *Target*: Yes for corporate ***hiring***.
	Explanation: MT made an error that compromises the source meaning due to lack of context, when translates ***contratação*** for ***hiring***. It is important to understand what is being hired. This information is given by previous information: **subscription plan**. Thus, instead of hiring, the conventionalised term **subscription** should be used.
Overly Literal: Word-for-word translations. This type of error translation denies access to the original text's meaning.	*Target*: O #PRS_ORG# não tem. *Source*: The #PRS_ORG# doesn't have it. *Target*: ***I see*** *Source*: ***Eu vejo***.
	Explanation: The MT does not recognise the expression in the source as a discourse marker, translating the expression word-for-word, creating an MT error due to an overly literal translation that compromises the source text meaning transfer.
Lexical Selection: Anomalous word choices considering the source sentence. Issue often associated to the translation of idiomatic expressions, metaphors, and even polysemous words.	*Source:* Fico aguardando o email. *Target*: I am waiting for the email. *Source:* ***Abraços*** *Target:* ***Embraces***
	Explanation: Explanation: Polysemous word in Pt with two translation possibilities in En: ***Hugs*** or ***Embraces***, plus poor intrasentential information creates inadequate MT translation for this context.
Named Entity (NE): NE errors caused by lack of context, which does not allow for the determination of the presence of a NE.	*Source:* O último foi há pouco e chama-se ***a última mulher***. *Target*: The last was today, shortly, and it is called ***the last woman***.
	Explanation: Poor contextual information in the sentence might be responsible for Named-Entities MT errors. This category is to be used only if there is not enough contextual information in the source sentence.
Register: Referring to language formality/ informality, speakers adapt their discourse according to the audience.	*Source*: How can I help ***you*** today. *Target:* Como posso ***te*** ajudar
	Explanation: Gender agreement: masculine ***cobrado***/ feminine ***cobrada*** beyond the sentence level. In the example, only by accessing previous information (context **afastada**) we are able to understand that we need the feminine translation **cobrada**.

Table 2: Contextual MT error categories with corresponding examples and explanations.

In the next section, we will focus on the implementation of the framework, putting the concepts into practice with an ecological dataset coupled with an inter-annotator agreement measure.

4. Framework application with a customer support chat dataset: methodology

From the outset, our research has followed a data-driven methodology, connecting theoretical concepts with real-world examples. For framework validation, we used a subset of the MAIA corpus (Farinha et al., 2022) from the WMT 2022 shared task on chat translation, with the following statistics:

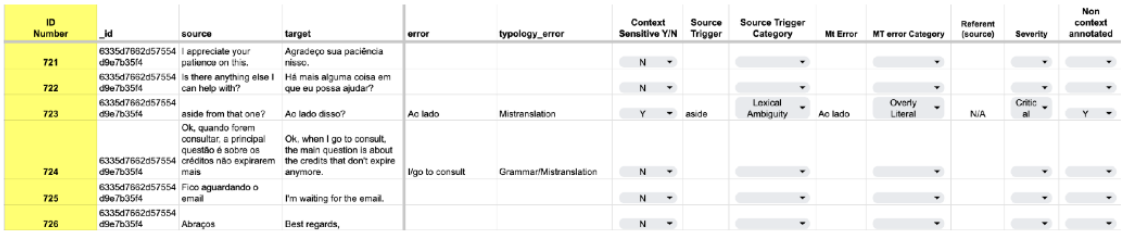
Maia Corpus	EN-PT/BR
Number of conversations	28
Number of agent segments	509
Number of customer segments	609
Total number of segments (customer and agent)	1168

Table 3: Statistics of the dataset used for context annotation, showing conversations, agent segments, customer segments, and total number of segments.

4.1. The annotation process

The annotation process was conducted by two experienced annotators who followed the context-aware annotation framework guidelines to perform a battery of tests set out in the form of a decision tree (both the tree and the instructions for its use are presented in the Annex). The annotators utilised a dedicated annotation scoreboard specifically designed for this task, where they simply filled in the required fields. The scoreboard

setting was designed through an iterative process to achieve a stripped-down format that would boost productivity. Figure 1 shows a snippet of the annotation scoreboard.



ID Number	_id	source	target	error	typology_error	Context Sensitive Y/N	Source Trigger	Source Trigger Category	Mt Error	MT error Category	Referent (source)	Severity	Non context annotated
721	6335d7662d57554d9e7b35f4	I appreciate your patience on this.	Agradeço sua paciência nisso.			N							
722	6335d7662d57554d9e7b35f4	Is there anything else I can help with?	Há mais alguma coisa em que eu possa ajudar?			N							
723	6335d7662d57554d9e7b35f4	aside from that one?	Ao lado disso?	Ao lado	Mistranslation	Y	aside	Lexical Ambiguity	Ao lado	Overly Literal	N/A	Critical	Y
724	6335d7662d57554d9e7b35f4	Ok, quando forem consultar, a principal questão é sobre os créditos não expirarem mais	Ok, when I go to consult, the main question is about the credits that don't expire anymore.	I/go to consult	Grammar/Mistranslation	N							
725	6335d7662d57554d9e7b35f4	Fico aguardando o email	I'm waiting for the email.			N							
726	6335d7662d57554d9e7b35f4	Abraços	Best regards,			N							

Figure 1: Annotation scoreboard used by Annotators 1 and 2 for inter-annotator agreement.

The complete definition for the full scorecard, by column, reads as follows:

Column Title	Definition
ID Number	Unique identifier for each sentence.
_id	Tracks dialogue boundaries for clarity.
Source/Target	Displays the source sentence and its MT output.
Error/typology_error	Preloaded MQM annotation from WMT 2022 for reference.
Context Sensitive (Y/N)	Flags whether the sentence depends on intersentential context.
Source Trigger & Category	Specifies the contextual MT error and its typology.
MT Error & Category	Specifies the contextual MT error and its typology.
Turn of referent	Measures the coreferential distance within the source dialogue.
Severity	Indicates the severity level of the MT error.
Non-context annotated	Marks whether the MT error was already annotated in WMT 2022 using MQM.

Table 4: Annotation scorecard description by column.

4.2. Inter-annotator agreement analysis methodology

To effectively evaluate our annotation framework, we compared the results from both professional annotators to assess inter-annotator agreement. We constructed a contingency table to compare annotations, using tags to capture total agreements, partial agreements, and full disagreements. The tags are as follows:

Category	Description
Tag (agr_no_contx)	Number of items with agreement on **no context** (True Positives).
Tag (agr_all_contx)	Number of items with agreement on **context** and **all annotation**.
Tag (disagr_contx_dep)	Disagreement on **contextual dependency**: One annotator finds context-dependent, the other context-independent.
Tag (disagr_src_trig)	Disagreement on **source trigger**.
Tag (disagr_error_cat)	Disagreement on **MT error category**.
Tag (disagr_error_sev)	Disagreement on **MT error severity.**
Tag (disgr_error_cat_plus_sev)	Disagreement on **both MT error categorisation and severity**.

Table 5: Annotation comparison tags in nominal values to be applied at the sentence level.

Applying the contingency table to each dialogue allowed for a comparative analysis of both annotators' decisions, offering a full perspective on each annotator's work. This process not only underscored the complexity of annotating context-dependent phenomena but also emphasised the benefits of an annotation framework designed to detect contextual MT errors and pinpoint their source-based triggers in a cause-effect relationship. The effectiveness of the proposed framework was evaluated using Cohen's κ metric (Cohen, 2013), which measures the degree of agreement between annotators beyond chance. The resulting κ score reflected a high level of annotation consistency, providing strong evidence of the framework's robustness and its capacity for generalisation across different annotators.

5. Inter-annotator agreement results

Each annotator's results are displayed in Table 6 for comparison. We start with an overall analysis of the contextual MT issues identified by each annotator, then examine each annotation dimension, reporting agreement and disagreement values, and conclude by measuring inter-annotator agreement by providing the full Cohen's κ metric scores for each category and the overall score.

Dependencies	Number of Sentences
N	1080
Y	87
Grand Total	**1167**

Dependencies	Number of Sentences
N	1019
Y	48
Grand Total	**1167**

Table 6: Context-sensitive sentence annotation results for Annotators 1 and 2. N represents the total number of sentences without any type of contextual trigger in the source sentence and, consequently, without MT errors linked with contextual triggers. Y represents the total number of sentences that have one or more contextual triggers in the source sentence and a corresponding MT error.

Focusing on the above results, it becomes apparent that annotation disagreements begin right at the surface level, with Annotator 1 flagging up 39 more context-dependent sentences than Annotator 2. In terms of percentages, Annotator 1 marked 7.45% of the test-set sentences as context-dependent, while Annotator 2 marked only 4.11%.

5.1. Cause-effect relationship between source triggers and MT errors

This section examines the cause-and-effect relationship between contextual source triggers and MT errors, identifying recurring patterns. The mapping visually links source triggers (vertical axis) to their corresponding contextual MT errors (colour-coded). Each bar represents the frequency of occurrences, showing how often a specific contextual trigger leads to an MT error and the most recurrent MT errors for that trigger.

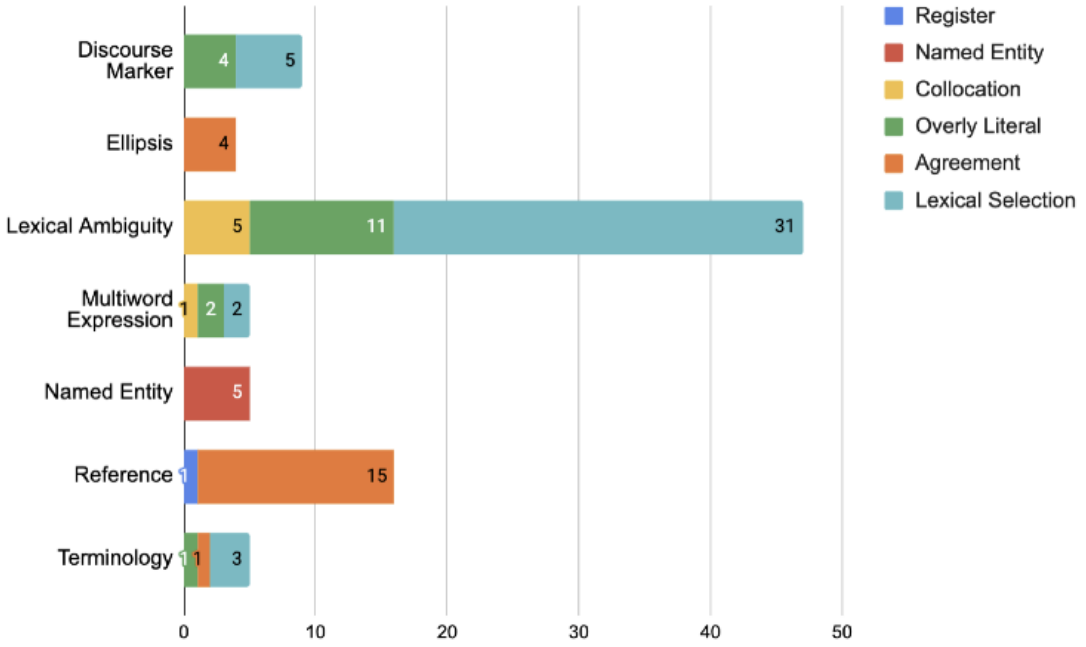


Figure 2: Contextual trigger and MT error mapping for Annotator 1.

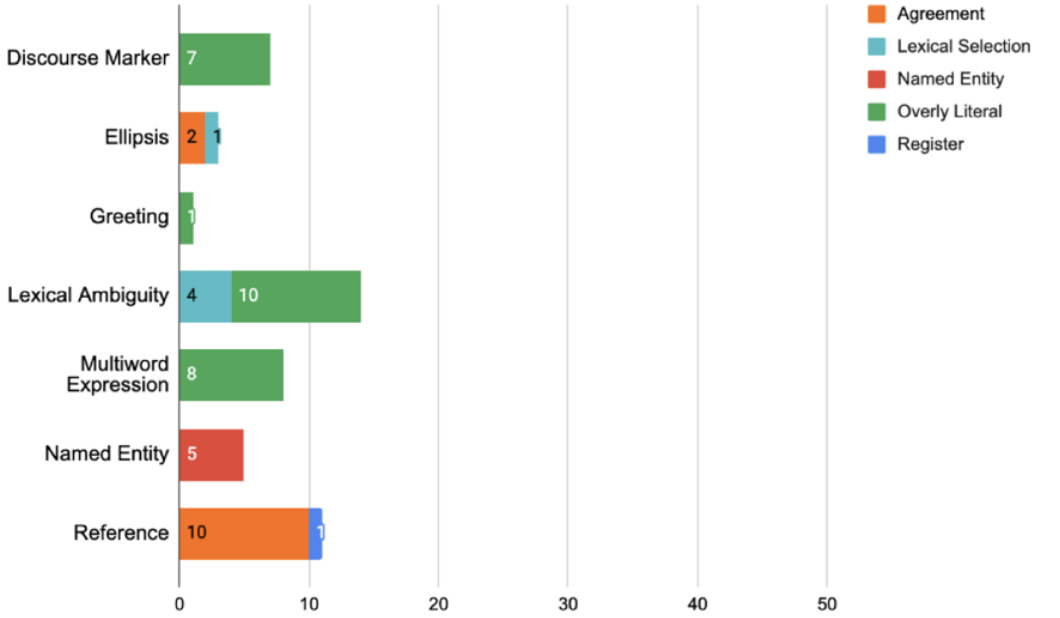


Figure 3: Contextual trigger and MT error mapping for Annotator 2.

Examining Figures 2 and 3, the source category *Lexical Ambiguity* stands out, in both sets of annotation results, as causing the most "contextual" MT errors. These results underscore that *Lexical Ambiguity* is a major challenge in the MT workflow, exposing the difficulty resolving such issues poses the system. The MT errors observed indicate a translation approach that fails to integrate broader contextual information, revealing a persistent limitation in the MT system.

The second most frequent cause of MT errors among the source trigger categories was *Reference*. A pattern observed in both annotators' work confirms that this category gives rise to two typical MT errors: *Agreement* and *Register* MT errors.

For the contextual trigger *Discourse Marker*, Annotator 1 identified nine occurrences, leading to a total of nine MT errors: four categorised as *Overly Literal* and five as *Lexical Selection*. In contrast, Annotator 2 identified only seven occurrences, all resulting in *Overly Literal* MT errors. These structures have a more pragmatic and discourse-organisational role, and do not have a clear lexical meaning when dealt with in isolation.

In the *Multiword Expression* category, Annotator 1 identified two *Lexical Selection* MT errors, two *Overly Literal* MT errors, and one *Collocation* MT error. Annotator 2, meanwhile, identified eight *Overly Literal* errors.

Annotator 1 deemed the *Ellipsis* and *Named Entities* categories to account for four and five MT errors respectively, with the former leading to four *Agreement* MT errors and the latter to five *Named Entities* MT errors. Annotator 2 also judged the *Named Entities* trigger to cause five *Named Entities* MT errors but found *Ellipsis* to give rise to three MT errors: two *Agreement* errors and one *Lexical Selection* error.

For the *Terminology* trigger, Annotator 1 identified five MT errors: three *Lexical Selection* errors, one *Overly Literal* error, and one *Agreement* error. In contrast, this

trigger category did not feature in Annotator 2's results. Finally, Annotator 2 identified one *Overly Literal* MT error as being linked to the translation of a *Greeting* trigger in the source. This trigger category did not feature in Annotator 1's results.

In general, and based on Figures 2 and 3 above, our findings consistently demonstrate that ambiguity poses the most significant challenge for the MT model. This reinforces the importance of contextual information and wider contextual windows in a document for proper disambiguation in an MT scenario.

5.2. Contextual MT error severity distribution

As noted previously, to properly assess the impact of context within an MT scenario, it is not just important to understand the number of MT errors linked to contextual phenomena; it is also important to account for the MT error severity distribution. The results of the MT error severity annotations performed, following an adapted version of Lommel et al. (2014), are shown in Figures 4 and 5.

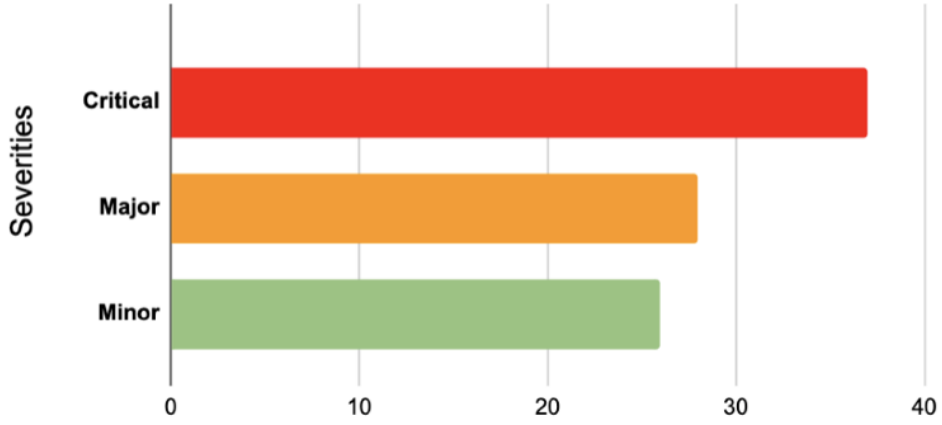


Figure 4: Contextual MT error severity distribution for Annotator 1.

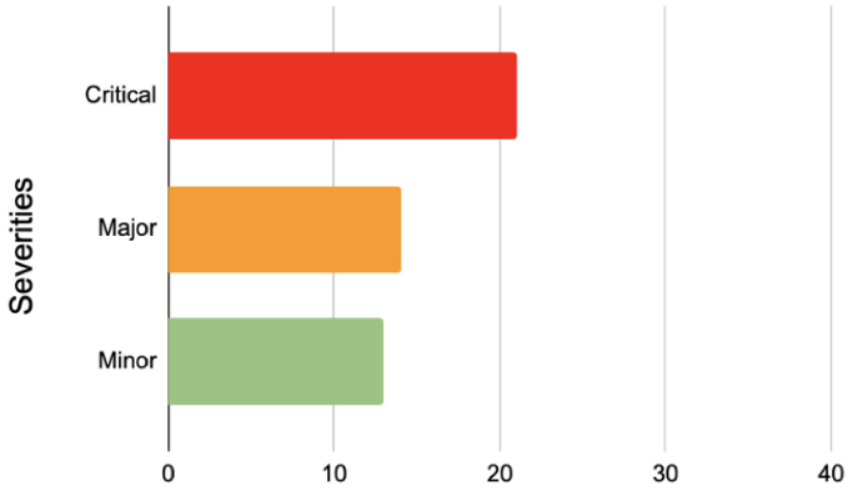


Figure 5: Contextual MT error severity distribution for Annotator 2.

As can be seen in the severity charts, the largest proportion of contextual triggers lead to critical MT errors, with Annotator 1 identifying 37 such instances (40.66% of the test set) and Annotator 2 identifying 21 (43.75%). In the case of the major MT error severity level, Annotator 1 tagged 28 MT errors (30.77%), while Annotator 2 tagged 14 (29.17%). Finally, Annotator 1 detected 26 minor MT errors (28.57%), compared to the 13 (27.08%) identified by Annotator 2.

Despite some imbalances in annotation values between the annotators, consistent severity patterns emerge. In terms of severity, critical MT errors represent the most significant portion of translation errors. This finding alone highlights the necessity of placing greater emphasis on contextual information in MT.

5.3. Inter-annotator dynamics: variability

Following the dialogue-level agreement analysis methodology described in Section 4.2, we ended up with the contingency table shown below, which represents the full inter-annotator schema.

Inter-annotator analysis	Total
Annotator agreement concerning the MT mistakes: No MT error output or the MT error has no contextual basis	1064
Annotator agreement concerning the source trigger identification	31
Annotator disagreement on contextual dependency in sentence	72
Annotator disagreement on source trigger category	2

Table 7: Descriptive annotation comparison table for the full test set.

As shown in the table above, the majority of the disagreements between annotators, accounting for 6.7% of the entire annotated dataset, concern determining whether an MT error is context-dependent or not. This result can be traced to the prototypical tendencies and trained behaviour of annotators who are accustomed to the traditional MT paradigm, focused on sentence-by-sentence analysis. Annotators typically approach error detection from a horizontal perspective, focusing on sentence-level comparisons between source and target. However, they are less familiar with our approach, which calls for a more vertical perspective that involves interpreting each sentence and MT error in the context of the entire document. The former annotation tendency is shaped by sentence-level MT evaluation, whilst our result highlights the shift in cognitive demands introduced by a context-aware annotation framework, rather than any shortcomings in the guidelines themselves, which provided clear instructions and illustrative examples. Next, we delve into a more detailed analysis of the annotated data.

As previously stated, through our inter-annotator agreement analysis methodology we identified the source trigger categories that caused greater uncertainty in annotations, leading to higher levels of disagreement in determining contextual dependency.

Disagreement between annotators in terms of context-dependent sentences, showing source trigger category annotations.	**Sum of Total**
Discourse Marker	10
Ellipsis	2
Lexical Ambiguity	37
Multiword Expression	12
Named Entity	2
Reference	5
Terminology	4
Grand Total	72

Table 8: Contextual source trigger categories tagged by Annotators 1 and 2 for tag disagr_contx_dep (disagreement on contextual dependency in sentences).

As previously mentioned, it is the categories that involve levels of ambiguity and require a certain degree of interpretation that led to more annotation disagreements at the sentence level, i.e. *Lexical Ambiguity, Multiword Expression*, and *Discourse Marker*. Furthermore, by consistently analysing the tag *disagr_contx_dep*, we captured every possible combination of categories (source trigger and MT error) contributing to disagreements on contextual dependency between the two annotators. This approach highlights the most problematic category pairings.

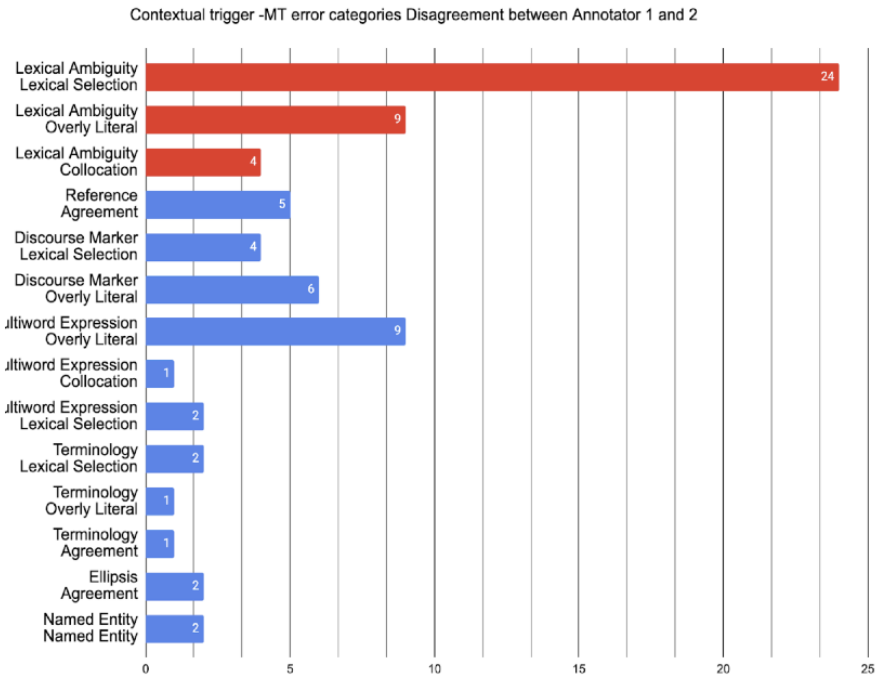


Figure 6: Source sentence dependency mismatches. Category pairings that caused the most annotation disagreements between Annotator 1 and Annotator 2 in terms of contextual dependency.

As shown in Figure 6, *Lexical Ambiguity* is the primary source of annotation discrepancies, frequently tied to *Lexical Selection* MT errors. Despite extensive guideline examples, this outcome was expected, as ambiguous lexical structures often lack precision, making analysis difficult and misclassification likely.

In contrast, categories like *Named Entities*, *Discourse Markers*, and *Register* are the source of minimal inter-annotator disagreement. The *Multiword Expressions* category is an exception, with higher disagreement due to the inherently ambiguous nature of such expressions (e.g. phrasal verbs). This reinforces how much of a challenge semantic ambiguities pose for both MT systems and human annotators, impacting interpretation and inter-annotator agreement.

5.4. Inter-annotator dynamics: agreement

Having previously presented our results regarding mismatches between annotators, we now report on the inter-annotator agreement (IAA) score obtained using the Cohen's κ metric.

Comparing annotated columns:	Cohen's K score
Context Sensitive (Y/N)	0.4410
Source Trigger	0.7902
Source Trigger Category	0.9182
MT Error	0.6553
MT Error Category	0.7968
Referent (source)	0.8031
Severity	0.7122
Average Agreement	**0.7310**

Table 9: IAA Cohen's κ scores by category and overall score.

Table 9 shows the IAA Cohen's κ scores for the different dimensions within the framework. According to Amidei et al. (2019: 347) and previous research work, it is safe to say that, for the most part, our IAA scores are set within the $0.6 < IAA \leq 0.8$ value range, representing a substantial correlation between the two annotators. Even in the case of the *Context Sensitive Y/N* category, where results are in the range of 0.4, this is considered satisfactory by Landis & Koch (1977), as reported by Amidei et al. (2019: 347). On this point, we have already discussed the reasons behind this score's disparity (for a more detailed discussion, see Section 5.3). The subsequent category scores specifically reflect the subset of data for which the annotators were in consensus that the MT error was context-based. These high values underscore the effectiveness of the

guidelines in offering a comprehensive framework and clear directions for annotators, validating the extensive effort invested in developing the detailed instructions and the decision tree (which can be found in the Annex). Overall, and considering the results reported, which give an overall agreement of 0.73 across the seven categorical columns, it can be confidently asserted that the proposed typology yielded highly positive outcomes, especially in the case of source trigger categorisation, underscored by a value of $0.8 < IAA \leq 1$, considered almost perfect.

We have developed an effective, robust annotation framework that has yielded promising results. It marks a significant advancement relative to existing frameworks like MQM by extending support for contextual phenomena. Used alongside traditional frameworks, it offers a more complete and impartial view of MT errors and performance, highlighting an MT system's capabilities more clearly. It also generates training and test datasets for developing context-aware MT models and context-sensitive automatic metrics.

We must acknowledge that the granularity and multidimensional nature of our framework significantly increase the cognitive load for annotators. Considering this, we are developing alternative annotation strategies to reduce the initial cognitive load on human annotators. For future steps, we will deploy LLMs as primary annotators, with humans serving as post-annotation reviewers and monitors of the process, in line with the concept of augmented translations (O'Brien, 2024). Our prompt testing has already achieved a strong balance between coverage and accuracy, revealing contextual MT errors previously missed by human annotators. However, further prompt refinement is required before scaling to a full annotation task across multiple annotators and language pairs.

6. Conclusion

With our research, we have shown the significance of context for MT, leading to new methods and architectures that shift towards document-level-oriented MT. We have delineated and defined the concept of context, which has been vital for understanding its different levels before tackling related issues.

We have described first attempts to overcome the limitations of QA models and frameworks through contextual error test suites, though these are scarce in terms of typology coverage and focus on commonly analysed domains. Instead, we propose an alternative annotation framework designed for document-level MT QA, covering the relatively untapped domain of customer support chat, closely related to spontaneous dialogues for contextual error analysis.

To assess MT quality from a contextual perspective, we used previously researched theoretical concepts on context to develop a comprehensive annotation ecosystem, including a full framework, annotation guidelines with a decision tree (available in the Annex), and a detailed analysis methodology.

Unlike most annotation frameworks and metrics, which prioritise one side of the translation process, our approach considers both source and target equally. This methodology identifies key connections and patterns between contextual MT errors and linguistic structures in the source document that trigger them. In our initial implementation, the framework's efficacy quickly became evident, helping identify MT shortcomings, particularly in handling lexical ambiguities. According to both annotators involved in the implementation, these lexical structures, which signal semantic variation based on context, account for most contextual MT errors; indeed, lexical ambiguities were the primary cause of disagreement between the two annotators. Additionally, the type of data chosen is crucial. We focused our research on MT of chat dialogues due to their distinctive traits and constraints, providing a fitting scenario for validating our framework.

Regarding IAA, the Cohen's κ score of 0.73 across the seven categorical columns shows strong agreement, confirming that our framework, guidelines, and decision tree generalise effectively and are reproducible across professionals. Based on these results, we can assert that our framework aligns with MT QA best practices. This lays a promising foundation for future frameworks offering a more inclusive approach to MT assessment, incorporating both vertical (full document) and horizontal (sentence-by-sentence) perspectives, creating the potential to provide more detailed training data for better-quality MT systems or context-sensitive automatic metrics.

In our next steps, we aim to refine the annotation guidelines to better support annotators. We will continuously test our framework by applying it to two customer support chat datasets — En–Ko and En–Pt/Br — translated by a fine-tuned LLM used as both a context-agnostic and context-aware MT model. These datasets, assessed with our framework, will allow us to evaluate the benefits of using LLMs as context-aware MT systems.

Additionally, to reduce the cognitive load on annotators, we are developing prompts that streamline the annotation process, using LLMs for initial tagging, with annotators serving as reviewers and validators. This approach creates a more efficient workflow, potentially lowering annotation costs while improving IAA.

7. Authors Contributions

Miguel Menezes: Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Writing – original draft.
Amin Farajian: Conceptualisation, Software, Data curation, Supervision.
Helena Moniz: Conceptualisation, Methodology, Validation, Formal analysis, Supervision, Project administration, Funding acquisition, Writing – review & editing.
João Graça: Supervisor

8. Funding

This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (i.e. the Center For Responsible AI); by Fundação para a Ciência e Tecnologia (FCT), through the project with reference "DOI:10.54499/UIDB/50021/2020"; by the Centro de Linguística da Universidade de Lisboa (CLUL), UID/214/2025; and through the FCT PhD grant with reference 2022.12091.BD.

8. Bibliography

Amidei, Jacopo; Piwek, Paul; Willis, Alistair (2019). Agreement is overrated: A plea for correlation to assess human evaluation reliability. In: Van Deemter, Kess; Lin, Chenghua; Takamura, Hiroya (eds,). In: van Deemter, Kess; Lin, Chenghua; Takamura, Hiroya (eds.). *Proceedings of the 12th International Conference on Natural Language Generation.* Association for Computational Linguistics, pp. 344–354. <https://aclanthology.org/W19-8642>. [Accessed: 20251217].

Bawden, Rachel (2018). *Going beyond the sentence: Contextual machine translation of dialogue* [Doctoral dissertation]. Université Paris-Saclay. Paris. <https://tel.archives-ouvertes.fr/tel-02066998>. [Accessed: 20251217].

Birner, Betty J. (2012). *Introduction to pragmatics*. Hoboken, NJ: John Wiley.

Bublitz, Wolfram (2011). Cohesion and coherence. In: Zienkowski, Jan; Östman, Jan-Ola; Verschueren, Jef (eds.). *Discursive Pragmatics. Handbook of Pragmatics Highlights.* Amsterdam; Philadelphia: John Benjamins, pp. 37–50. <https://doi.org/10.1075/hoph.8>. [Accessed: 20251217].

Cai, Xiaoyu; Xiong, Deyi (2020). A test suite for evaluating discourse phenomena in document-level neural machine translation. In: Liu, Qun; Xiong, Deyi; Ge, Shili; Zhang, Xiaojun (eds.). *Proceedings of the Second International Workshop on Discourse Processing*. Association for Computational Linguistics, pp. 13–17. <10.18653/v1/2020.iwdp-1.3>. [Accessed: 20251217].

Castilho, Sheila; Doherty, Stephen; Gaspari, Federico; Moorkens, Joss (2018). Approaches to human and machine translation quality assessment. In: Moorkens, Joss; Castilho, Sheila; Gaspari, Federico; Doherty, Stephen (eds.). *Translation Quality Assessment: From Principles to Practice.* Cham: Springer, pp. 9–38.

Castilho, Sheila; Cavalheiro Camargo, João Luiz; Menezes, Miguel; Way, Andy (2021). DELA corpus: A document-level corpus annotated with context-related issues. In: Barrault, Loic; *et al.* (eds.). *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 1–12. <https://aclanthology.org/2021.wmt-1.63/>. [Accessed: 20251217].

Cohen, Jacob (1988). *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Escribe, Marie (2019). Human evaluation of neural machine translation: The case of deep learning. In: Temnikova, Irina.; Orasan, Constantin.; Corpas Pastor, Gloria.; Mitkov, Ruslan (eds.). *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*. Association for Computational Linguistics, pp. 36–46. <https://aclanthology.org/W19-8705>. [Accessed: 20251217].

Fang, Qiong (2018). A study of the impact of translation ecosystem on the translator from the perspective of restriction factors. *IOP Conference Series: Materials Science and Engineering*, v. 452, n. 3, 032020. <https://doi.org/10.1088/1757-899X/452/3/032020>. [Accessed: 20251217].

Farinha, Ana C.; Farajian, M. Amin; Buchicchio, Marco; Fernandes, Patrick; De Souza, José G. C.; Moniz, Helena; Martins, André F. T. (2022). Findings of the WMT 2022 shared task on chat translation. In: Koehn, Philipp; *et al.* (eds.). *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Association for Computational Linguistics, pp. 724–743. <https://aclanthology.org/2022.wmt-1.72>. [Accessed: 20251217].

Freitag, Markus; Rei, Ricardo; Mathur, Nitika; Lo, Chi-Kiang; Craig, Stewart; Foster, George; Bojar, Ondřej (2021). Results of the WMT21 metrics shared task: evaluating metrics with expert-based human evaluations on TED and News Domain. In: Barrault, Loic; *et al.* (eds.). *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, pp. 733–774. <https://aclanthology.org/2021.wmt-1.74>. [Accessed: 20251217].

Garvin, David A. (1984). What does "quality" really mean? *Sloan Management Review*, v. 25, n. 1, pp. 25–43.

Grice, H. Paul (1991). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Hassan, Hany; Aue, Anthony; Chen, Chang; Chowdhary, Vishal; Clark, Jonathan; Federmann, Christian; *et al.* (2018). *Achieving human parity on automatic Chinese-to-English news translation.* ArXiv:1803.05567. <https://doi.org/10.48550/arXiv.1803.05567>. [Accessed: 20251217].

Habermas, Jürgen (1979). *Communication and the evolution of society*. Boston: Beacon Press.

Halliday, Michael A. K. (1989). *Language, context and text.* Geelong: Deakin University Press.

Horn, Laurence R.; Ward, Gregory L. (eds.) (2004). *The handbook of pragmatics*. Oxford: Wiley.

Horton, William S. (2012). Shared knowledge, mutual understanding and meaning negotiation. In: Hans-Jörg Schmid (ed.). Cognitive Pragmatics. Berlin; Boston: De Gruyter Mouton, pp. 375–398.

Jin, Lifeng; He, Jie; May, Jonathan; Ma, Xuezhe (2023). *Challenges in context-aware neural machine translation*. arXiv:2305.13751. <https://doi.org/10.48550/arXiv.2305.13751>. [Accessed: 20251217].

Jwalapuram, Prathyusha; Rychalska, Barbara; Joty, Shafiq; Basaj, Dominik (2021). DiP benchmark tests. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2004.14607>. [Accessed: 20251217].

Koby, Geoffrey S.; Fields, Paul; Hague, Daryl R.; Lommel, Arle; Melby, Alan (2014). Defining translation quality. *Revista Tradumàtica: tecnologies de la traducció,* n. 12, pp. 413–420. <https://doi.org/10.5565/rev/tradumatica.76>. [Accessed: 20251217].

Läubli, Samuel; Sennrich, Rico; Volk, Martin (2018). *Has machine translation achieved human parity? A case for document-level evaluation*. ArXiv:1808.07048. <https://doi.org/10.48550/arXiv.1808.07048>. [Accessed: 20251217].

Läubli, Samuel; Castilho, Sheila; Neubig, Graham; Sennrich, Rico; Shen, Qinlan; Toral, Antonio (2020). A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research,* v. 67, pp. 653–672. <https://doi.org/10.1613/jair.1.11371>. [Accessed: 20251217].

Lommel, Arle; Uszkoreit, Hans; Burchardt, Aljoscha (2014). Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció,* n. 12, pp. 455–463. <https://ddd.uab.cat/pub/tradumatica/tradumatica_a2014n12/tradumatica_a2014n12p455.pdf>. [Accessed: 20251217].

Lommel, Arle; Gladkoff, Serge; Melby, Alan; Wright, Sue Ellen; Strandvik, Ingegerd; Gasova, Kristyna; Nenadic, Goran (2024). *The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control.* arXiv:2405.16969. <https://arxiv.org/abs/2405.16969>. [Accessed: 20251217].

Malinowski, Bronisław (2000). The problem of meaning in primitive languages. In: Lucy Burke; Tony Crowley; Alan Girvin (eds.). *The Routledge Language and Cultural Theory Reader.* London; New York Routledge, pp. 386–395. [Accessed: 20251217].

Menezes, Miguel; Farajian, M. Amin; Moniz, Helena; Varelas Graça, João (2023). A Context-Aware Annotation Framework for Customer Support Live Chat Machine Translation. In: Utiyama, Masao; Wang, Rui (eds.). *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, Macau SAR China.* Asia-Pacific Association for Machine Translation, pp. 286–297. <https://aclanthology.org/2023.mtsummit-research.24/>. [Accessed: 20251217].

Müller, Mathias; Rios, Annette; Voita, Elena; Sennrich, Rico (2018). *A large-scale test set for pronoun translation*. ArXiv:1810.02268. <https://arxiv.org/abs/1810.02268>. [Accessed: 20251217].

Nord, Christiane (2014). *Translating as a purposeful activity*. London: Routledge.

O'Brien, Sharon (2023). Human-centered augmented translation. *Perspectives*, v. 32, n. 3), pp. 391–406. <https://doi.org/10.1080/0907676X.2023.2247423>. [Accessed: 20251217].

Papineni, Kishore; Roukos, Salim; Ward, Todd; Zhu, Wei-Jing (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In: Pierre, Isabelle; Charniak, Eugene; Lin, Dekang (eds.). *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 311–318. <https://aclanthology.org/P02-1040>. [Accessed: 20251217].

Petrick, Fabian; Herold, Christian; Petrushkov, Pavel; Khadivi, Siavash; Ney, Hermann (2023). *Document-level language models for machine translation*. ArXiv:2310.12303. <https://arxiv.org/abs/2310.12303>. [Accessed: 20251217].

Post, Matt; Junczys-Dowmunt, Marcin (2023). *Escaping the sentence-level paradigm.* ArXiv:2304.12959. <https://doi.org/10.48550/arXiv.2304.12959>. [Accessed: 20251217].

Rei, Ricardo; Stewart, Craig; Farinha, Ana C.; Lavie, Alon (2020). *COMET: A Neural Framework for Mt Evaluation.* ArXiv:2009.09025. <https://arxiv.org/abs/2009.09025>. [Accessed: 20251217].

Shen, Lihong (2012). Context and text. *Theory and Practice in Language Studies*, v. 2, n. 12, pp. 2663–2669. <https://www.academypublication.com/issues/past/tpls/vol02/12/28.pdf>. [Accessed: 20251217].

Silverman, Hugh J. (1986). What is textuality? Part II. *Phenomenology + Pedagogy,* v. 4, n. 1, pp. 54–61. <https://doi.org/10.29173/pandp15010>. [Accessed: 20251217].

Stalnaker, Robert (2002). Common ground. *Linguistics and Philosophy*, v. 25, n. 5–6, pp. 701-721. <https://doi.org/10.1023/A:1020867916902>. [Accessed: 20251217].

Tierney, Robert J.; Mosenthal, James H. (1983). Cohesion and textual coherence. *Research in the Teaching of English,* v. 17, n. 3, pp. 215-229. <https://www.jstor.org/stable/40170955>. [Accessed: 20251217].

Toral, Antonio; Castilho, Sheila; Hu, Ke; Way, Andy (2018). *Attaining the unattainable? Reassessing claims of human parity in neural machine translation*. ArXiv:1808.10432. <https://doi.org/10.48550/arXiv.1808.10432>. [Accessed: 20251217].

Vermeer, Hans J. (1978). *Ein Rahmen für eine allgemeine Translationstheorie*. Heidelberg: Groos.

Vernikos, Giorgos; Thompson, Brian; Mathur, Prashant; Federico, Marcello (2022). Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. ArXiv:2209.13654. <https://doi.org/10.48550/arXiv.2209.13654>. [Accessed: 20251217].

Von Wright, Georg Henrik (1981). Explanation and understanding of action. *Revue internationale de philosophie,* v 35, n. 135, pp. 127–142. <https://www.jstor.org/stable/23945379>. [Accessed: 20251217].

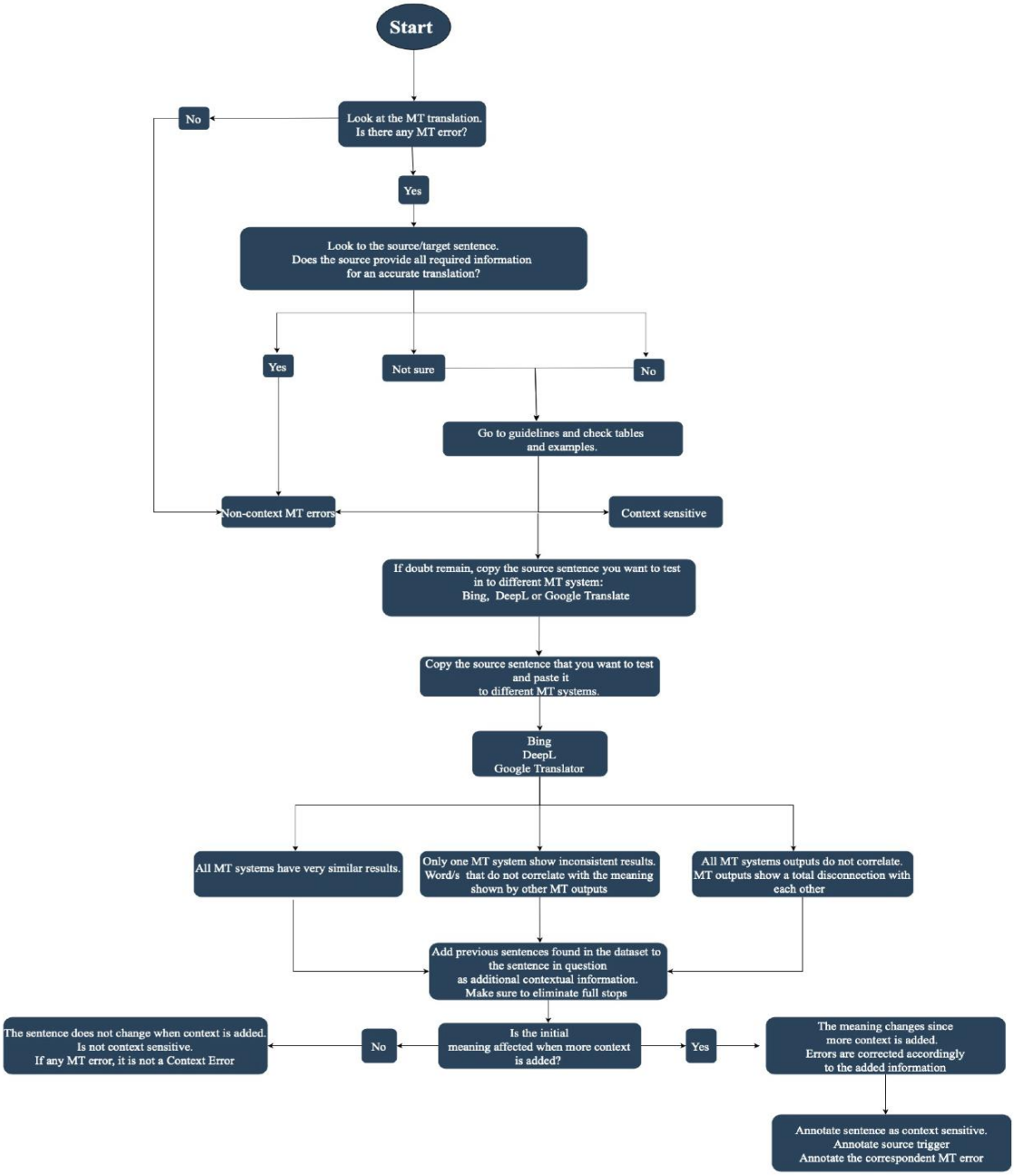
Wicks, Rachel; Post, Matt (2023). Identifying context-dependent translations for Evaluation Set Production. In: Koehn, Philipp; *et al.* (eds.). *Proceedings of the Eighth Conference on Machine Translation (WMT), December 6-7, 2023.* Association for Computational Linguistics, pp. 452-467. <https://aclanthology.org/2023.wmt-1.42/>. [Accessed: 20251217].

Wittgenstein, Ludwig (1958). *Philosophical investigations*. Oxford: Blackwell.

Yin, Kexin; Fernandes, Patrick; Pruthi, Danish; Chaudhary, Aditi; Martins, André F. T.; Neubig, Graham (2021). *Do context-aware translation models pay the right attention?* ArXiv:2105.06977. <https://doi.org/10.48550/arXiv.2105.06977>. [Accessed: 20251217].

Annex

Decision tree for contextual dependency assessment.



Instructions for use:

Step one: The annotator should contrast source and target sentences.

Step two: The annotator should proceed through the decision tree by first posing the following question:

Is there any MT error?

If the answer is No, there is nothing to annotate, as there is no visible contextual trigger. The annotator should follow the decision tree until reaching the Non-Context MT Errors option and then move on to the analysis of the next sentence.

If, on the other hand, the answer is Yes, the annotator should continue following the decision tree until reaching the next step.

Step three: Examine the source and target sentences. Does the source provide all the information required for an accurate translation?

If Yes, the annotator should continue along the decision tree and proceed directly to Non-Context MT Errors, then move on to the next sentence analysis.

If the annotator is Not sure or answers No, they should refer to the guidelines for more detailed information, carefully analysing both the definitions and the examples provided. These resources will support effective decision-making and help resolve any uncertainties.

Let us now move to the practical application of the decision tree using the following example taken from the analysed dataset:

Source:

You will need Wi-Fi access, the password of your #PRS_ORG# account to log in back

Target:

Você precisará de acesso Wi-Fi, a senha da sua conta #PRS_ORG# para se inscrever para trás.

This process should be interactive and may be applied as many times as necessary. Additionally, the annotator should remain at each decision point until every question has been satisfactorily answered.

Question:

Does the source provide all the information for an accurate translation even if the target presents translation inaccuracies?

Answer:

Yes. One can easily understand that the expression “log in back” links to password and password links to Wi-Fi access. All essential required for an accurate translation are contained within the sentence itself. In this case, the annotator reaches the Non-Context MT Error category and can confidently proceed to the analysis of the next sentence.

If the MT output contains errors that undermine the annotator’s confidence regarding contextual dependencies, the same step should be repeated by asking the following questions:

Question:

(i) Does the MT contain errors? If so, can they be linked to any source-trigger category?

(ii) Could the error involve, for example, an ambiguous word that requires information provided before or after the sentence to enable accurate translation?

Answer:

(i) Yes, the MT output contains an error.

(ii) However, although the MT translation of *"log in back"* is incorrect, the expression itself is not ambiguous in translational terms. All the information required for an accurate translation is present within the sentence boundaries (i.e. intrasentential information).

It is worth noting that even a seemingly simple word such as *log* can be polysemous or homonymous, with meanings such as:

"a part of the trunk or a large branch of a tree that has fallen or been cut off".

"an official record of events during the voyage of a ship or aircraft".

However, it is the intrasentential context—that is, the surrounding words within the same sentence—that constrains the intended meaning of log and eliminates ambiguity. This contextual restriction results in a sentence whose meaning can be fully determined without reliance on information beyond the sentence boundaries.

If the annotator determines that the MT errors are non-contextual, as in the example analysed above, they may confidently proceed to the analysis of the next sentence. If, however, the annotator identifies the MT error as being triggered by contextual factors, they should follow the decision-tree path leading to the Context-Sensitive option. In such cases, the sentence is considered context-dependent while also exhibiting an MT error. Consequently, both the contextual trigger in the source sentence and the MT error must be annotated in accordance with the framework.

From the above analysis, the following conclusion can be drawn: the sentence functions as a self-contained unit and does not require intersentential contextual information to be correctly understood or translated. It contains all the information necessary for accurate MT processing and is therefore independent of external contextual input. In this case, the annotator should not annotate any contextual error and should assign an N (No) value under the Context Sensitive Y/N column in the scorecard. If uncertainty remains, the annotator should continue following the appropriate decision-tree path.

Let us now shift our focus to the practical application of the remaining section of the decision tree by presenting another example that has proven challenging in the identification of intersentential contextual dependencies.

Source:

I see!

Target:

Eu vejo!

Step One:

The annotator should copy the source sentence and submit it to multiple MT systems. All MT outputs should then be compared. We recommend the use of the following

commercial MT systems: Microsoft MT, DeepL MT, and Google MT. Annotators are advised to use multiple systems because they differ in proficiency and behaviour. This comparison facilitates the identification of whether any MT system adjusts or corrects a lexical structure when additional context is provided, thereby altering the translation's meaning.

Applying Step One to the example above yields the following results.

I see!	WMT MT system: Eu vejo!
	DeepL: Estou a ver!
	Google: Eu vejo!
	Bing: Eu vejo!

Step Two:

Based on the MT outputs, annotators may follow one of three paths:

- Path 1: All MT systems produce very similar results.
- Path 2: One MT system produces a result (word or words) that does not align with the others.
- Path 3: All MT systems produce markedly different outputs, suggesting a disconnect between source meaning and translation.

When applied to the "I see" example, Step Two produces the following observations:

(i) All MT systems translate the sentence as referring to the physical ability to perceive with the eyes.

(ii) The DeepL system shows a slight nuance, interpreting "I see" as a possible interjection meaning "I understand."

(iii) Given these results, further investigation is required to fully interpret this nuanced translation. Regardless, annotators using the decision tree must, without exception, proceed to Step Three.

Step three:

This step is mandatory and essential for testing contextual dependency, regardless of the path selected in Step Two. Here, previous sentences are incorporated to test whether the meaning of the sentence under analysis changes when additional context is provided. Adding extra-sentential context involves including preceding sentences from the same dialogue.

When applying this strategy, particular care must be taken. To avoid sentence breaks that could compromise the validity of the test, the additional context must be integrated compositionally, as though forming part of a single coherent sentence:

Source:
Just to confirm, does #PRS_ORG# already have a corporate account and corporate plans set up with #PRS_ORG#? The #PRS_ORG# does not have it, I see!

Target:
Só para confirmar, #PRS_ORG# já possui uma conta corporativa e planos corporativos configurados com #PRS_ORG#? O #PRS_ORG# não tem, entendo!

As demonstrated in the example above, the manner in which contextual information is added is crucial. The translation of "I see" was adjusted to entendo, reflecting the intended meaning within the dialogue. Notably, the initial MT error was corrected once contextual information was introduced, resulting in an accurate and contextually appropriate translation.

This process should be repeated across multiple MT systems to allow for comparison and to determine whether different systems adapt their outputs in response to the added context, potentially resolving the original MT error.

Step Four:

After completing this process, the annotator addresses the following question:

Question:

Does the addition of context affect the meaning?

Answer:

Yes. The inclusion of contextual information results in an adjusted MT output, indicating contextual dependency.

At this point, the annotator can reach the following conclusion: the sentence in question is context-sensitive. The lexical structure "I see" in the source sentence should therefore be identified as a contextual trigger responsible for the context-dependent MT error "Eu vejo".