Bringing Context into MT Evaluation: Translator Training Insights from the Classroom

Sheila Castilho *

Posted Date: 8 August 2025

doi: 10.20944/preprints202508.0597.v1

Keywords: machine translation; large language models; AI-based translations; human evaluation; context-aware evaluation; translation students; translator training

 doi:10.20944/preprints202508.0597.v1

Article

Bringing Context into MT Evaluation: Translator Training Insights from the Classroom

Sheila Castilho

School of Applied Language and Intercultural Studies/The ADAPT Centre, Dublin City University; sheila.castilho@dcu.ie

Abstract

The role of technology in translator training has become increasingly significant as machine translation (MT) evolves at a rapid pace. Beyond practical tool usage, training must now prepare students to engage critically with MT outputs and understand the socio-technical dimensions of translation. Traditional sentence-level MT evaluation, often conducted on isolated segments, can overlook discourse-level errors or produce misleadingly high scores for sentences that appear correct in isolation but are inaccurate within the broader discourse. Document-level MT evaluation has emerged as an approach that offers a more accurate perspective by accounting for context. This article presents the integration of context-aware MT evaluation into an MA-level translation module, in which students conducted a structured exercise comparing sentence-level and document-level methodologies, supported by reflective reporting. The aim was to familiarise students with context-aware evaluation techniques, expose the limitations of single-sentence evaluation, and foster a more nuanced understanding of translation quality. This study provides methodological insights for incorporating MT evaluation training into translator education and highlights how such exercises can develop critical awareness of MT's contextual limitations. It also offers a framework for supporting students in building the analytical skills needed to evaluate MT output in professional and research settings.

Keywords: machine translation; large language models; AI-based translations; human evaluation; context-aware evaluation; translation students; translator training

1. Introduction

The role of technology in translator training has been a subject of study since technology became prominent in the wider field of translation. Teaching translation technology to translators is not only important for practical purposes but also for understanding the historical and socio-technical aspects of translation [1].

Research on translator training regarding technical competencies has been extensive. Doherty et al. [2] presented a project focused on teaching Statistical Machine Translation (SMT) to post-graduate student translators. The authors concluded that when the role of the translator is made more active rather than passive, students and translators are empowered by having a greater familiarity and understanding from the developers' perspective.

Doherty and Kenny [3] argued for the implementation of a new translator-oriented approach to teaching machine translation (MT) in the classroom. At the time, SMT was not widely taught, and when it was, the role of translators in the SMT workflow was constructed in a limiting way (p. 277), due to the tendency "to see post-editing as the only role for translators (who then morph into post-editors) in such workflows" (p. 287). They describe an experiment on implementing SMT in a postgraduate course, where translators are seen as empowered users of this technology, rather than serving solely as bastions of translation quality at the end of the translation process. Their results (described in Doherty and Kenny [3]) showed a significant increase observed in "SMT self-efficacy and ample evidence of enhanced critical appreciation among students of the technologies involved" (p. 310). The significant

 doi:10.20944/preprints202508.0597.v1

advances of SMT at the time also lead to proposing that post-editing and NMT should be incorporated across the full curriculum [4].

With the emergence of Neural MT (NMT) systems and the improvement in translation quality, particularly for languages where SMT was lacking [5], students have eagerly embraced MT and rely on it extensively [6]. Thus, it is crucial to educate translators in training about the capabilities and limitations of the available MT systems.

As evident, there is a widespread acknowledgment that education in translation technology must adapt to new technologies and the evolving nature of the translation sector. Therefore, this article advocates for the importance of teaching MT evaluation from a perspective that integrates context-aware evaluation, which has gained increasing attention in MT research with the advent of NMT systems and Large Language Models (LLMs). While the role of context has long been recognised in translation studies, its systematic application in MT evaluation has only recently emerged as a critical area of research, driven by the need for more rigorous assessment methodologies [7]. To explore this issue, we pose the following questions:

Q1: How can a context-aware MT evaluation exercise be effectively implemented in translator training?
Q2: What is the impact of context-aware MT evaluation on students' ability to assess translation quality compared to sentence-level evaluation?

The next section presents works that have attempted to teach MT evaluation in the classroom.

2. Related Work

2.1. MT Evaluation in the Classroom

Human evaluation of MT has been fundamental to the field since its inception. Assessing the quality of MT systems involves the use of either automatic or human evaluation methods, each possessing its own strengths and weaknesses [8].

Scholars have emphasised the significance of incorporating MT education into translator training, enabling them to "recommend and use the most appropriate evaluation methods in a given scenario" [3] [p.285]. A comprehensive survey of MT competencies conducted by Gaspari et al. [9] [p.333] highlighted "a strong need for improvement in quality assessment methods, tools, and training" due to the increased integration of technology into translation curricula. The survey also indicated that post-editing was the most common type of evaluation performed.

With the paradigm shift from SMT to NMT models, the importance of evaluation was even more pronounced as researchers sought to assess improvements in fluency and adequacy, as well as new types of errors introduced by NMT [5]. Studies from the SMT-to-NMT transition period show that comparative evaluations played a key role in helping students understand MT quality [10].

Moorkens [11] reported a practical translation evaluation exercise in the classroom, based on the evaluation outlined in [5], where undergrads and PhD students compared the quality of SMT and NMT systems via post-editing, adequacy, and error typology. The exercise enabled students to identify errors in MT output and comprehend the complexities of evaluation, particularly when NMT produced fluent but inaccurate translations. Results of this exercise showed that most of the students found they gained hands-on experience of a new MT paradigm and were positive towards the technology, and expressed interest in working with NMT in future. Moorkens reported that this experiment enabled students to be empowered rather than fear the technology, and, moreover, in subsequent classes based on this exercise, students discussed scenarios where use of NMT may be appropriate (perishable texts) and highly inappropriate (transcreation, literary texts, jobs that involve regulatory compliance). While SMT has now been largely phased out, these early evaluations laid the groundwork for integrating MT evaluation training into translator education, equipping students with the skills to assess different modern MT systems.

 doi:10.20944/preprints202508.0597.v1

Various other studies have implemented MT in the classroom, including post-editing training [12], task-based approaches (e.g., terminology, word order, passive voice) [10], post-editing training and its impact on creativity [13], and post-editing and pricing strategies [14], among others.

In light of this "dynamic technological environment with standard practices regularly overridden by new and updated tools and technologies," teaching translation technology poses challenges [11] [p.375]. However, building upon Kenny [1]'s argument that "curricula have to be updated to integrate technologies such as neural MT" and that "translation students and practitioners have to remain alert to the possibility of AI bias in MT" (508), we argue here that teaching new methodologies for MT evaluation is as essential as teaching about the technology itself, particularly given the current stage of advancement in the translation field, with generative AI being used not only for translation but also for evaluation [15].

The following section provides an overview of the evaluation methods that consider context to evaluate not only the quality but also potential harms of translation outputs. While there has been some research in this area, it remains limited, and to date, no studies have explored integrating these insights into translator training.

2.2. The Role of Context in MT Evaluation

With the rise of NMT systems and the emergence of LLMs, assessing these systems has become increasingly challenging due to their high fluency, tendency to generate plausible yet inaccurate translations (hallucinations), and the necessity of discourse-level evaluation to capture coherence across longer texts. Differentiating between MT output and human translation has proven difficult using current human evaluation methods [16,17], as 'segment-level scores may not provide a sufficiently comprehensive framework to address the intricate challenges introduced by context-aware translations' [7]. Context in MT evaluation can take multiple forms, including intersentential dependencies, world knowledge and external information, and multimodal input [18]. As a result, there is a growing demand for assessment approaches that better account for these dimensions of context. One such approach, which has gained increasing attention in the MT community, is document-level human evaluation, as it allows for a more thorough assessment of suprasentential coherence and discourse-level phenomena [19].

Several studies have attempted to determine the amount of context necessary for translators [20,21], as well as various methodologies for incorporating context into MT evaluation [19,22,23].[1]

Castilho ([19,22]) tested different methodologies in order to add context to MT evaluation. She found that when sentences are evaluated in a single-sentence random set-up translators tend to have higher inter-annotator agreement, compared to set-ups where sentences are given in context. However, there is a high number of misevaluation cases found in the single-sentence set-ups. Moreover, the context-evaluation enabled translators to find harmful errors and made them more confident in the evaluation. The author notes that the high IAA scores in the single sentence setup are due to the fact that translators "tend to accept the translation when adequacy is ambiguous but the translation is correct, especially if it is fluent" (p.42). Subsequently, both the Conference for Machine Translation (WMT) and The Metrics Shared Task have integrated context evaluation into their practices, employing various methodologies since 2021.

It is evident that the field of MT evaluation is dynamic, and the most effective methodologies for conducting context-aware human evaluations are yet to be identified. However, it is vital to show translators in training the fast-changing field, as they need to be able to identify the strengths, weaknesses, and possible harms of these advanced systems with a more rigorous evaluation. While there has been some research into context-aware MT evaluation, no work has yet explored how to bring these insights into the context of translator training. In light of this, the exercise described in this article attempted to teach master's students about different methodologies for MT evaluation, based on previous work by Castilho [19,22].

[1] For a comprehensive survey on the role of context in MT evaluation, see Castilho and Knowles [24].

 doi:10.20944/preprints202508.0597.v1

3. Context of this Study and Educational Setting

This evaluation task was administered to two cohorts of master's students at Dublin City University as part of the Translation Technology module. This module is a mandatory component of both the MA in Translation Studies[2] and the MSc in Translation Technology[3] programmes.

3.1. The Translation Technology Module

As mentioned earlier, this task was conducted during the Translation Technology lab, a module designed to familiarize students with specialized technologies utilized in the translation field. The primary objective of this module is to provide students with hands-on experience and promote critical reflection on the utility of these technologies. The specific learning objectives of the module include:

- Understanding the principles underlying contemporary translation technologies
- Proficiency in using commercial translation memory tools
- Enhancing texts for machine translation or human translation facilitated by a translation memory tool
- Grasping the principles of contemporary machine translation
- Critically assessing contemporary translation technologies and their outputs
- Making informed decisions regarding the utilization of these technologies while considering associated risks
- Understanding current professional, socio-technical, and ethical issues and trends related to technology in the profession, research, and industry/market.

The Translation Technology module carries a weight of 10 ECTS (European Credit Transfer System) credits, aligning with the standards set by the Irish higher education sector. This equates to one-sixth of the overall taught master's program.

3.1.1. Lectures

The translation technology module comprises eleven two-hour lectures and eleven one-hour labs, with labs scheduled after the lectures. During the initial four weeks of the module, the focus is on Computer-Assisted Translation (CAT) tools, covering topics such as the evolution of the profession, technical aspects of tools, and socio-technical issues in translation technology. Correspondingly, the first four labs are dedicated to instructing students on the utilization of two CAT tools, elucidating their features in alignment with the lectures.

Beginning from week 5, the curriculum shifts towards MT, commencing with a brief historical overview followed by exploration of different paradigms (rule-based, statistical, and neural MT, and Large Language Models since 2023). Subsequently, the focus transitions to MT evaluation, encompassing human evaluation, automatic evaluation, and post-editing. The lab sessions in week 6 introduce students to the concept of human evaluation, particularly emphasising fluency, adequacy, and error annotation. Weeks 8 and 9 centre on learning to evaluate and train a NMT system online, and utilising online platforms with automatic metrics, respectively.

The context-aware evaluation task was assigned to students during the week 10 lab session, subsequent to an extensive lecture on human evaluation.

3.1.2. Student Demographics

Approximately one-third of the students enrolled in the program are of Irish nationality, with English as their native language or, to a lesser extent, Irish. The remaining cohort primarily comprises students from other European Union countries where English is not the primary language, as well as individuals from regions including North and South America, Asia, and the Middle East. Their ages typically range from 22 to 50, with an average age of approximately 26. The majority of students hold undergraduate degrees in languages or combinations with other disciplines, while a minority possess

[2] For more information, refer to: https://www.dcu.ie/courses/postgraduate/salis/ma-translation-studies
[3] For more information, refer to: https://www.dcu.ie/courses/postgraduate/salis/msc-translation-technology

 doi:10.20944/preprints202508.0597.v1

undergraduate degrees in translation studies. Upon admission, students generally exhibit proficiency in standard office applications, though they typically lack experience in computer programming. Additionally, the majority have limited exposure to translation technologies, with only a handful having encountered technologies with basic interfaces such as online MT interfaces. In 2022, the module had 30 students enrolled, increasing to 38 students in 2023.

4. Methodological Setup: Sentence vs. Document Evaluation Lab

As previously mentioned, the lab comparing sentence and document-level evaluation was conducted in week 10, following students' exposure to human and automatic evaluation metrics.

The objective of the lab was to illustrate to students the biases inherent in sentence-level evaluations and the disparities in inter-annotator agreement, as discussed in [22]. Specifically, the lab aimed to expose students to the potential limitations of sentence-level MT evaluation by comparing it to document-level evaluation, which takes into account contextual information from surrounding sentences. This comparison was designed to help students understand how context-aware evaluation can lead to more accurate and nuanced assessments of MT output.

4.1. Metrics

As previously mentioned, students had already received lectures on human evaluation metrics, thus they were familiar with the definitions. Nevertheless, each metric was defined in the lab instructions and within the spreadsheet.

A. Adequacy

Adequacy assessment pertains to the conveyed meaning of the translated text. It is measured using a Likert scale, with respondents answering the question: "How much of the meaning expressed in the source appears in the translation?"

1. None of it
2. Little of it
3. Most of it
4. All of it

B. Fluency

Fluency assessment concerns the grammatical aspects of the translated text. It is measured using a Likert scale, with respondents answering the question: "How fluent is the translation?"

1. No fluency
2. Little fluency
3. Near native
4. Native

C. Error Annotation

Error annotation comprises a simple taxonomy with four error types:

- Mistranslation: The target content does not accurately represent the source content.
- Untranslated: Content that should have been translated has been left untranslated.
- Word Form: There is a problem in the form of a word (e.g., the English target text has "becomed" instead of "became").
- Word Order: The word order is incorrect (syntax) (e.g., "The house beautiful is" instead of "The house is beautiful").
- No errors: The target content is flawless.

D. Ranking

Ranking evaluation requires the translator to select the best version of a translation (MT1 or MT2). If the translator believes that both versions exhibit the same quality, they can choose 'Both'.

 doi:10.20944/preprints202508.0597.v1

4.2. Materials

The lab materials provided detailed instructions on conducting the evaluation, along with a spreadsheet featuring five distinct tabs:

- Tab 1: Sentence-level evaluation of fluency, adequacy, and error (Figure 1)
- Tab 2: Sentence-level evaluation of ranking (Figure 2)
- Tab 3: Inter-annotator agreement (Figure 3)
- Tab 4: Document-level evaluation of fluency, adequacy, and error (Figure 4)
- Tab 5: Document-level evaluation of ranking (Figure 5)

4.3. Corpus

The evaluation tasks were based on the document-level DELA corpus [25], a test suite specifically designed to assess document-level machine translation quality. The corpus contains English-language source texts annotated for six types of context-related phenomena: lexical ambiguity, grammatical gender, grammatical number, terminology, ellipsis, and reference. These annotations identify instances where correct interpretation or translation of a segment depends on document-level context rather than sentence-level cues alone. The annotations were created in the English source texts, not in the translations, with the intention of highlighting context-sensitive challenges that are likely to cause errors in translation, particularly for MT systems that operate at the sentence level. The DELA corpus has previously been used to evaluate whether context-aware MT systems could resolve such issues more effectively than sentence-based systems. Students were asked to evaluate existing MT outputs in their chosen language pair, but the source of difficulty remained the English-language context. It is important to note that while not every context issue necessarily triggers an error in all language pairs, the selected examples were known (from previous MT evaluations) to be problematic for at least some MT systems and language pairs. Thus, the purpose was not to guarantee an error per sentence, but rather to prompt students to examine context carefully and recognise when context might be necessary for proper evaluation.

For the 2022 evaluation, six full documents were selected from the DELA corpus: four from the Review domain and two from the News domain. From these, 21 context-annotated sentences were selected for the sentence-level evaluation (i.e. presented out of context), while the document-level evaluation included the full texts with the same 21 sentences highlighted. This dual presentation enabled comparison between evaluations performed with and without context. However, based on student feedback and observed time constraints, the 2023 evaluation adopted a lighter version. Only three shorter Review domain documents were used, and the number of context-annotated sentences selected for sentence-level evaluation was reduced to 15.

4.4. Translation Systems

In the 2022 lab session, students were instructed to use an online MT system of their preference to translate the source text into their target language. Specifically, one MT system served as their MT1 for the evaluation of fluency, adequacy, and error (tabs 1 and 4 – see section 4.2), while another MT system served as their MT2 for the ranking evaluation (tabs 2 and 5– see section 4.2), as the ranking evaluation needed the use of two distinct MT systems.

However, for the 2023 lab, in order to streamline the process and save time, the MT1 columns were pre-populated with translations generated by Google Translate for all target languages across all tabs. Subsequently, students were tasked with selecting an online MT system or a large language model of their choice to serve as their MT2. Notably, some students opted to use ChatGPT[4] as their MT2.

[4] https://chat.openai.com/

 doi:10.20944/preprints202508.0597.v1

4.5. The Lab

The lab was designed to raise students' awareness of how context can affect translation evaluation, and consisted of two sequential tasks:

- Task 1: Sentence-Level Evaluation – Students evaluated a selection of individual sentences for fluency, adequacy, error identification, and ranking, without access to surrounding context. They then collaborated with a peer working in the same language pair to compute inter-annotator agreement (IAA) scores.
- Task 2: Document-Level Evaluation – Students were presented with the full documents from which the same sentences had been extracted. They re-evaluated the same sentences, now situated in their original context, and again computed IAA and compared both sets of scores (from Task 1 and Task 2).

Students were instructed to review the lab instructions. They had the option to download the files and work with them locally or use them on the cloud platform. They were explicitly advised not to discuss their evaluations with peers in the same language pair, in order to minimise scoring influence and preserve the independence of IAA calculations.

We note that re-evaluating the same sentences in Task 2 may introduce a bias stemming from prior exposure. However, this design choice was intentional and pedagogically motivated. By working with the same sentences across both tasks, students were encouraged to reflect critically on how their judgments shifted with access to context. Rather than viewing the second evaluation as "correcting" the first, the aim was to stimulate meta-cognitive insight into the challenges of sentence-level evaluation and the role of document-level information in interpreting MT output. This design enabled students not only to quantify differences in IAA and quality scores across tasks but also to engage in qualitative reflection on the types of context-sensitive errors that are likely to go unnoticed in sentence-level evaluation. In post-lab debriefing and written reflections, several students explicitly commented on their surprise at how context changed their interpretation of MT quality — an intended outcome of the exercise.

4.5.1. Task 1: Sentence-level Evaluation

To carry out Task 1, students were asked to open the first tab, "1 - SENT: Adequacy, Fluency and Errors", to conduct the human evaluation of a text at sentence level in terms of adequacy, fluency, and error annotation (see section **??**). All assessments were performed via a drop-down menu interface (see Figure 1).

	A	B	C	D	E	F	G	H
1			ADEQUACY	FLUENCY	ERROR			
2	Source	MT	How much of the meaning expressed in the source appears in the translation?	How fluent is the translation?	Select the errors you found in the MT output. You can select multiple error types and same error type more than once.			
3								
4	I am going to have my son try to turn the legs the other way and hope it helps.							
5	The massive organ fills the space and adds to the romance!		4. All of it					
6	They will test your will.		3. Most of it					
7	Assembled, the it is gorgeous, durable and very easy to use.		2. Little of it					
8	They are AWFUL.		1. None of it					
9	It was in my opinion rather unremarkable other than its fame.							
10	They will test your endurance.							
11	It is pretty.							
12	If you get it unassembled, God speed.							
13	I think my husband put it together backwards because it barely rocks.							
14	They will test your mental stability.							
	I was amazed by the detail and the sheer feeling it gave off!							

Figure 1. Tab 1: Sentence-level evaluation of fluency, adequacy, and error. A drop-down menu was available for adequacy, fluency (Likert 1-4) and error (4 categories).The MT column was filled in by the 2022 cohort and pre-filled by the lecturer for the 2023 cohort.

 doi:10.20944/preprints202508.0597.v1

Upon completion of tab 1, students moved to tab 2 "2 - SENT: Ranking" to perform the ranking evaluation on the same set of sentences, selecting the best translation among MT1, MT2, or Both, using the drop-down menu (see Figure 2). Subsequently, the results of Task 1 were discussed in class.

	A	B	C	D
1	Source	MT1	MT2	BEST TRANSLATION Choose your preferred translation
2				
3				
4	I am going to have my son try to turn the legs the other way and hope it helps.			
5	The massive organ fills the space and adds to the romance!			BOTH
6	They will test your will.			MT1 MT2
7	Assembled, the it is gorgeous, durable and very easy to use.			
8	They are AWFUL.			
9	It was in my opinion rather unremarkable other than its fame.			
10	They will test your endurance.			
11	It is pretty.			
12	If you get it unassembled, God speed.			
13	I think my husband put it together backwards because it barely rocks.			
14	They will test your mental stability.			
15	I was amazed by the detail and the sheer feeling it gave off!			
16	It is more of an olive green than a mint green.			
17	And they will test your sailor vocabulary.			
18				

Figure 2. Tab 2: Sentence-level ranking evaluation was conducted using a drop-down menu featuring options for 'MT1', 'MT2', and 'BOTH'. The MT1 column was filled in by the 2022 cohort and pre-filled by the lecturer for the 2023 cohort. The MT2 column was filled in by the students themselves.

Following the conclusion of Task 1, students moved to Tab 3 "IAA: Adequacy&Fluency" to compute the IAA for adequacy, fluency, and error. They were instructed to collaborate with another student sharing the same language pair (and using the same MT system, for the 2022 cohort). The tab already contained Cohen's Kappa formula [**?**], only requiring students to input their scores as Translator 1 (T1) and their classmate's scores as Translator 2 (T2) (see Figure 3). They then analysed their scores, and the results were discussed in class.

	A	B	C	D	E	F	G	H	I	J
1	T1	T2								
2	2	2					T2			total
3	3	3				1	2	3	4	
4	4	4			1	1	0	0	0	1
5	4	4		T1	2	0	2	0	0	2
6	4	3			3	0	0	3	1	4
7	3	4			4	0	0	1	3	4
8	3	3		total		1	2	4	4	11
9	1	1								
10	2	2								
11	3	3				1	2	3	4	
12	4	4			agreement	1	2	3	3	9
13					by chance	0.090909091	0.363636364	1.454545455	1.454545455	3.363636364
14										
15					KAPPA	0.738095238				
16										
17					1	2	3	4		
18				No:	2	4	8	8		
19				Per:	9.090909091	18.18181818	36.36363636	36.36363636	100	
20										

Figure 3. Inter-annotator agreement calculation. Students filled in Column A and B with their own scores and the scores provided by a colleague.

4.5.2. Task 2: Document-Level Evaluation

Finally, students proceeded to Task 2, opening the Tab 4 "1 - DOC: Adequacy, Fluency and Errors". They were directed to assess the sentences highlighted (in yellow) within each text/document but were required to read the full document to understand the context in which these target sentences appeared. Students evaluated adequacy, fluency, and identified error types using the options provided

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2025 doi:10.20944/preprints202508.0597.v1

in the drop-down menus (see Figure 4).[5] Students evaluated adequacy, fluency, and identified error types using the options provided in the drop-down menu (see Figure 4).

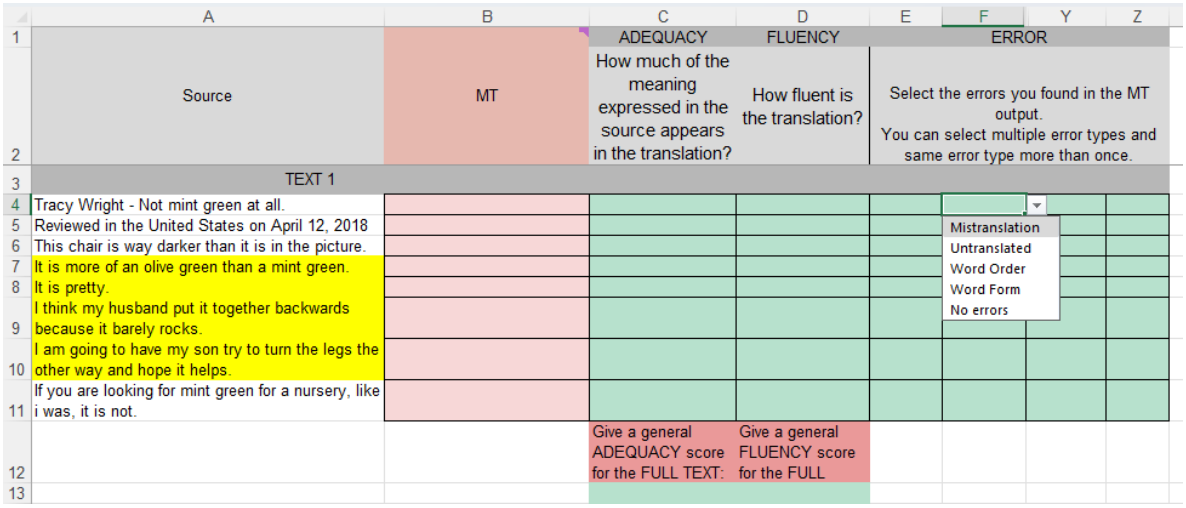

Figure 4. Tab 4: Document-level evaluation of fluency, adequacy, and error. A drop-down menu was available for adequacy, fluency (Likert 1-4) and error (4 categories).The MT column was filled in by the 2022 cohort and pre-filled by the lecturer for the 2023 cohort. Note that the highlighted sentences were the ones students evaluated in the sentence-level set-up.

Next, students were instructed to add the MT2 output (the same version used during the sentence-level evaluation) to Tab 5, "2 - DOC: Ranking," and to evaluate the translation at the document level, this time focusing on ranking MT1 and MT2 in context (see Figure 5). As in the sentence-level ranking task, MT2 was used solely for comparative ranking purposes and was not independently evaluated for adequacy, fluency, or error type.

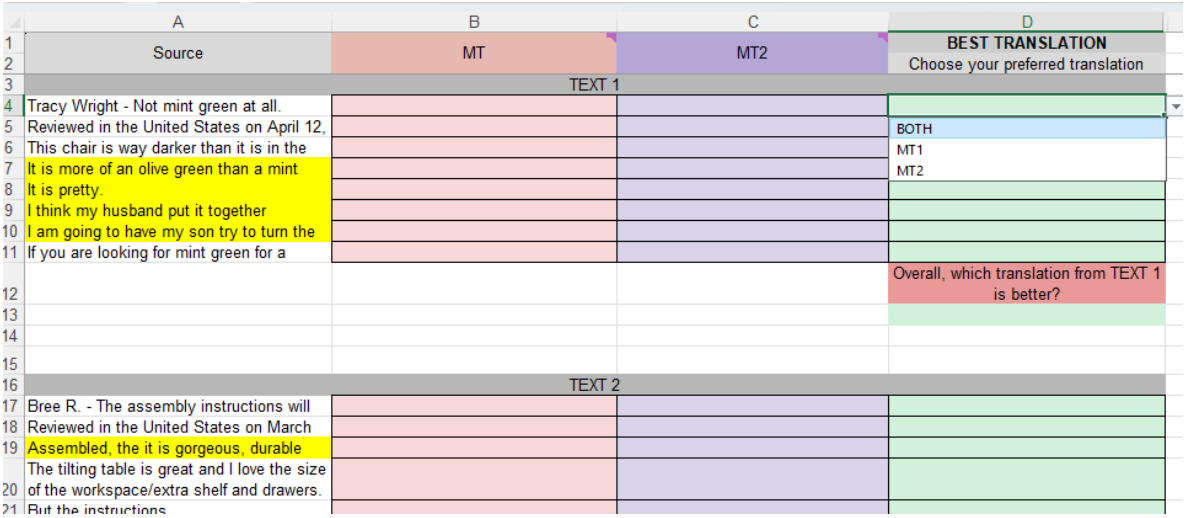

Figure 5. Tab 5: Document-level ranking evaluation was conducted using a drop-down menu featuring options for 'MT1', 'MT2', and 'BOTH'. The MT1 column was filled in by the 2022 cohort and pre-filled by the lecturer for the 2023 cohort. The MT2 column was filled in by the students themselves. Note that the highlighted sentences were the ones students evaluated in the sentence-level set-up.

[5] Due to time constraints during the lab session, students were asked to evaluate only the highlighted sentences rather than every sentence in the document, so they would have sufficient time to complete the IAA task. However, those who finished early were encouraged to assess the remaining sentences within the same documents.

 doi:10.20944/preprints202508.0597.v1

Upon completion of the evaluation, they were instructed to return to Tab 3, the IAA Tab and input their scores for the document-level evaluation as T1, while partnering with a classmate as T2 (see Figure 3).

Subsequently, students analysed their scores and discussed their findings, highlighting the disparities between evaluating at the sentence level and the document level. This discussion occurred both in pairs and in class with the lecturer.

It is worth mentioning that the document-level tabs included an option for students to assign one single score for the entire translated documents (as shown in Figures 4 and 5). However, due to time constraints, this task could not be completed during the lab session by most students. Students were advised to complete this task at home or in the lab if they finished the evaluation early. They were required to assign one single score for adequacy, fluency, and ranking to each full text, and to reflect on the rationale behind their scoring decisions. This was intended to illustrate the complexity and subjectivity involved in assigning a single quality score to an entire document, particularly when translation quality varies across segments. As reported in Castilho [22]), assigning one score per full document tends to result in lower inter-annotator agreement for adequacy, ranking, and error identification when compared to sentence-level scoring, and sentence-in-context scoring. This is likely because full documents often contain a mix of "very good," "reasonably good," and "poorly translated" sentences, which makes the overall adequacy harder to pin down consistently. Translators may understand the document reasonably well overall, yet weigh different parts of it differently when forming a judgment, leading to divergent interpretations of what score best represents the text as a whole. This exercise was designed to help students recognise the challenges of holistic document-level assessment and to reflect critically on how translation quality is unevenly distributed across a text, affecting their ability to make consistent evaluative decisions.

5. Insights from Student Reflections

As part of the lab exercise, students were asked to submit a short Reflective Report (up to 500 words), articulating their thoughts on the two evaluation approaches: sentence-level and document-level assessment. To support their reflection, a structured questionnaire was provided, consisting of the following eight prompts:

- What were your expectations in terms of translation quality before doing this exercise?
- Were the results as you had expected?
- How did you approach evaluating the MT output(s) quality with single random sentences in contrast to evaluating sentences in context (with the full texts)?
- As a translator, how did you feel assessing the translation quality with the two different methodologies (single sentence vs sentence in context) in terms of:
  a. effort
  b. how easy it was to understand the meaning of the source
  c. how easy was to recognise translation errors.
  Give examples.
- Did you have any cases of misevaluation? If so, explain why you think they happened and give examples (source, target, gloss translation).
- Were the IAA results (if any) as you expected? Comment on the differences.
- With which methodology were you most satisfied and confident with the evaluation you provided? Why?
- Will this change how you use MT?

We note that these were intended purely as guidance, and therefore, students were not required to respond to each question individually, but rather to use them as a starting point for their own critical reflections. What follows is a synthesis of the most relevant and revealing insights that emerged across the reflective reports.

 doi:10.20944/preprints202508.0597.v1

5.1. Expectations

The purpose of this opening question was to prompt students to reflect on their prior knowledge and understanding of MT quality.

The responses revealed that the majority of students held specific expectations regarding the performance of the MT systems they were utilising prior to the evaluation. Interestingly, students with language pairs known to pose fewer challenges for MT, such as romance languages, tended to have higher expectations of the system's quality. Conversely, students dealing with language pairs recognised as more challenging for MT, such as low-resource languages, generally held lower expectations. It was noteworthy to observe that several students, having previously experimented with multiple MT systems, opted to explore entirely different systems that they had not previously encountered, such as eTranslation and GPT.

5.2. Approach to sentence vs document-level evaluation

This question was aimed to make students reflect on their approaches to MT evaluation. Notably, this was a key focal point emphasised by the lecturer during the lab, encouraging students to consider potential adjustments or shifts in their evaluation strategies.

Several students indicated that their approach to sentence-level evaluation primarily centred on "focusing on fluency" (with some mentioning "grammar") of individual translated sentences, without needing to "think of the context", as ensuring adequacy would require additional contextual information. Students noted that during sentence-level evaluation, they only checked for "obvious errors" to determine their scores:

> *"I tackled each sentence individually, concentrating on their linguistic correctness"*

> *"When assessing the translation quality, I found it is harder to evaluate the accuracy of a single random sentence than that of a sentence in full context, thus I was inclined to focus more on the assessment of fluency. For example, the word 'they' in the sentence 'They will test your will' can be translated as 他们or 它们depending on the context, the former one generally refers to human beings, while the latter refers to non-human things, animals or abstract concepts. In the absence of context, it can be assumed that both translations are correct."*

In contrast, students mentioned that in the document-level approach, they were able to assess both fluency and adequacy, as the full context facilitated a more comprehensive evaluation. Some students reported reading the entire source text before commencing their assessment, while others preferred to evaluate sentence by sentence, occasionally revisiting earlier judgments when encountering new contextual information. Rather than viewing this as a source of bias, the design of the task intentionally encouraged students to re-examine and potentially revise their initial sentence-level evaluations in light of new contextual insights. This reflective process was central to the pedagogical aim: to demonstrate how evaluative decisions are context-sensitive and to prompt students to critically assess their own evaluation strategies.

5.3. Perception of Difficulty

This question intended to encourage students to contemplate the level of effort involved in various forms of human evaluation of MT across different metrics. Additionally, it encouraged students to consider how the two different methodologies (sentence-level vs. document-level) may facilitate or hinder the accurate assessment of errors, and how the ability to comprehend the source text can affect the evaluation.

Effort: In terms of perceived effort, there was a divide in numbers of students who thought the sentence-level evaluation required more effort than the document-level evaluation. Interestingly, students who mentioned that they struggled "focusing on fluency" and/or "grammar" only (see Section 4.2) to evaluate the random sentences, described that this methodology required more effort, while students who tended to accept the source text's ambiguity more easily stated that more effort

 doi:10.20944/preprints202508.0597.v1

were required in the document-level evaluation, since they had more text to read and make sure the translation was consistent.

> *"I felt less strict on my interpretation and evaluation and more indulgent with imprecision and ambiguities* [for the sentence-level evaluation]– *and this exactly because I did not have the context. My perception of the task was positive: the work seemed to flow more easily and to demand small cognitive effort. Evaluating sentences in context in its turn, was a much more intellectually demanding task, requiring rereadings and a heavier load of interpretation that made my approach to the sentences more subjective and the categorisation of mistakes more demanding. However, it gave me the lacking context and allowed me to perceive mistakes that were invisible before"*

Understanding the Source and Identifying Errors: All students mentioned that errors were much easier to be identified when sentences were evaluated in context. The majority of students mentioned that the source sentences that contained some kind of ambiguity could only be fully understood in the document-level evaluation. Some of the most cited examples were sentences with ambiguous pronouns that could be translated in either gender (it, they), Named Entities (Chase Center), and lexical ambiguity examples (legs, juice). Interestingly, although acknowledging that it was easier to identify errors in contextualised sentences, some students mentioned that certain awkward sentences did not need the context to be evaluated:

> *"some of the mistakes were very obvious and in that sense it did not really matter if it was a single sentence or not, but for more complex or subtle mistakes having context surrounding the sentence that was being analysed did help to point them out"*

5.4. Misevaluation

The goal of this question was to guide students to critically reflect on instances of misevaluation, considering how they may have occurred and what measures could have prevented them. Given the diverse language pairs among our students (See section 2, the responses encompassed a wide range of perspectives. A considerable number of students noted that scores assigned to sentences in the sentence-level evaluation tended to be higher, even when the sentences contained ambiguous elements:

> *"The initial confusion was caused by the lack of context which prevented me from identifying the gender. Nevertheless, there was no ground to mark down the MTs' outputs as incorrect as there was no background information to prove this"*

Moreover, students reported instances where they adjusted the scores assigned for a translation in the sentence-level evaluation upon evaluating the same sentence with context, realising that the translation was, in fact, incorrect:

> *"The presence of the context influenced some of my decisions made on a sentence-based level. E.g., it was only after I had read the whole text, did I understand the full meaning of 'I am going to have my son try to turn the legs the other way and hope it helps' and, as a result, was able to adjust my final evaluation"*

The following are examples of misevaluation stemmed from ambiguity cases that were pointed out in some students' reflections, including:

Ambiguous pronouns (gender/reference):

- **It** is pretty.
- **They** will test your will.
- **They** are AWFUL.
- Bring **it**!

Lexical ambiguity:

- I am going to have my son try to turn the **legs** the other way and hope it helps.

 doi:10.20944/preprints202508.0597.v1

- The massive **organ** fills the space and adds to the romance!
- **Warriors** hope fans' return to Chase Center gives team **juice**
- **Bring it**!

Additionally, students with Japanese and Chinese language backgrounds observed that some of the types of errors included in the typology used for the task were not applicable or did not occur in their language pairs. After exploring the full MQM taxonomy [26], they identified instances in the MT outputs that they felt were context-related errors, but which did not clearly align with any of the categories provided.[6] This response demonstrated their critical engagement beyond the core task, as they pursued independent inquiry into the limitations of the error categories and their applicability across diverse linguistic systems.

5.5. Inter-Annotator Agreement - IAA

The question about the IAA was designed to prompt students to compare the agreement results between themselves when conducting evaluations at the sentence-level and the document-level.

The majority of the students were able to pair up with a colleague who shared the same language pair (and the same MT systems) to calculate IAA at the sentence-level. However, due to time constraints, only a few were able to compare the IAA at the document-level.

Students who calculated IAA solely at the sentence-level commented on the discrepancies in evaluation between themselves and their colleague, particularly regarding the choice of scores and error identification. Furthermore, they acknowledged the impact of the chosen MT system and the challenges posed by their language pairs, indicating an awareness of these factors.

Those who computed IAA at both the sentence-level and document-level reflected on the differences in IAA scores between the two evaluation tasks. The majority observed that IAA scores at the sentence-level were higher compared to those at the document level.

> *"These results may be explained by the fact that translators do not stress the aspect of gender as much on a sentence-level, but they become duly aware of that aspect once the evaluation is carried out on a document-level"*

5.6. Satisfaction

The question about which methodology the students were satisfied the most, and more confident with the evaluation they provided, prompted them to consider the advantages and drawbacks of each approach. This allowed them to reflect both as translators, assessing their confidence in assigning scores, and from the perspective of researchers or project managers designing translation evaluations.

A few students mentioned that they were confident when assessing the translations on the sentence-level due to the reduced effort required to consider the full context:

> *"I was more confident evaluating translations on sentence-level because it has less emphasis on the gender aspect. The other methodology* [document-level] *required more attentiveness from the translator to catch translation errors that manifest as minor gender inflection issues, which can be daunting especially for big projects"*

However, the majority of students indicated greater satisfaction with the document-level evaluation methodology and expressed more confidence in their scores using this approach:

> *"I was confident and satisfied with document level because I knew exactly what was expected of me. I was confident at sentence level until I got to document level and context was brought in"*

> *"the methodology I am most satisfied with is the sentence-in-context assessment with the adequacy metric. This approach allows for a comprehensive understanding of the source text and a clearer evaluation of the referential relationship. It also reduces inaccurate evaluations"*

[6] While students discussed these observations in general terms int heir reflective reports, they did not provide specific textual examples. As such, their reflections could not be analysed in detail, but still point to important cross-linguistic considerations in MT evaluation.

 doi:10.20944/preprints202508.0597.v1

> *"Overall, I preferred the evaluation at the document level, as I was able to understand the context and be more critical of the quality of the MT output, whereas I was not confident when evaluating the output at the sentence-level since I had no context to understand fully the source sentences or to judge if there was a gender error"*

This aligns with one of the core objectives of the exercise, to prompt students to experience firsthand how context influences both the reliability and confidence of translation evaluation. By contrasting the two methodologies, the task was designed not only to highlight the contextual blind spots of MT systems but also to make students aware of how these blind spots affect their own judgments as evaluators. In general, there was a connection between satisfaction and the perceived effort required for each method, with students appreciating the greater reliability of document-level evaluation despite its higher cognitive demands.

5.7. Future use of MT

Several reflective reports included insightful comments on how this activity might alter their approach to using MT. Some students noted that the exercise increased their awareness of MT's limitations and the importance of defining quality criteria in evaluations. They expressed a newfound caution and understanding of the need for clear evaluation frameworks:

> *"I never considered the term "quality" to be so difficult to operationalise. It is now clear that when trying to measure quality one should always define what is perceived as "high" or "low" in their evaluation, as well as state the framework it takes place in"*

Others expressed a more nuanced view of MT, acknowledging its limitations and the necessity of human evaluation for handling nuances in grammar:

> *"This will change my thought on how to use MT. As translators, we should be more aware how to link different MT systems and evaluation methods to different situations"*

> *"In sum, this type of evaluation already makes me perceive MT with gentler eyes and understand it has limitations since it cannot always deal with nuances related to grammar errors properly, which is rather up for human evaluation"*

> *"I am more aware of the errors that are prevalent in machine translation regarding fluency and accuracy problems, and it will be interesting to see the progression of Irish in various translation engines in the future"*

Interestingly, many students highlighted the importance of providing context to MT systems for better outputs. This was an interesting finding, especially for the 2022 cohort, as at that time, apart from big MT systems (like Google and DeepL), most of the open MT systems were not able to recognise any context dependencies. That means if students pasted a single sentence or the full text in the MT systems, there would be no changes in the translation. However, for some MT systems, the change between pasting single sentences and the full documents did happen:

> *"This will change the way I use machine translation. When I use machine translation, I will put the entire text into the online machine translation system instead of copying and pasting sentence by sentence into the machine translation system"*

> *"In order to get better outputs from the online translation engine, I suppose it is better to input all the sentences with context rather than single random sentences extracted from a text, since the engine can recognize the context and then generate better translations"*

> *"I think the way I use MT in the future will change in the way that I will try to give the MT more context instead of just one sentence, since I saw that it can change the translation"*

> *"This highlights the importance of providing a MT engine with context to ensure a higher possibility of receiving accurate output"*

 doi:10.20944/preprints202508.0597.v1

This finding prompted further experimentation and led to the publication of a paper with the students on how open MT systems handle context.

6. Insights from the Exercise

This exercise provided valuable insights into the nuances of evaluating MT output at different levels of granularity. The aim of this lab was to show students the variance in translation quality resulting from evaluations conducted at the sentence and document levels. Students were able to learn how to rigorously assess translations from EN into their respective languages, using the state-of-the-art human metrics consisting of adequacy, fluency, error and ranking.

The selection of the DELA corpus proved to be a good fit for this evaluation, as the context-related issues annotated in the English source texts (such as reference, grammatical gender, and lexical ambiguity) were relevant to many of the language pairs used by the students. While these issues did not manifest in the same way across all languages, they were sufficiently cross-cutting to stimulate meaningful reflection on context-sensitive translation challenges. Additionally, the methodological decision to highlight specific sentences for sentence-level evaluation helped structure the task effectively: it enabled students to begin with a focused sentence-level assessment and later revisit those same examples within the full document context, thereby supporting a direct and pedagogically valuable comparison.

The accessibility and utility of the spreadsheet-based evaluation tool extended beyond the lab itself, with many students incorporating it into subsequent assignments and dissertations.

However, certain limitations of the exercise were noted. Some students appeared to misconstrue the lab's primary objective, focusing on comparing the quality of the two MT systems used (MT1 and MT2), rather than comparing the two evaluation methodologies (sentence-level vs. document-level). In hindsight, this misunderstanding is understandable: while MT2 was included solely as a reference point for ranking, it was not intended to be evaluated independently for adequacy, fluency, or errors - a distinction that may not have been sufficiently emphasised. This highlights the need to communicate the pedagogical aims more explicitly in future iterations of the lab, particularly regarding the role of secondary MT outputs in the evaluation process.

Moreover, emphasising how different MT systems handle context is essential to helping students understand the source of certain translation errors. Some systems struggle with context dependencies, particularly in areas such as reference, gender, or lexical ambiguity. Others may generate more accurate output when provided with full-document input, reducing the number of misevaluations students encounter due to missing contextual cues. In the current lab setup, the same MT outputs (typically generated from sentence-level input) were used in both evaluation tasks. However, this may have introduced bias, as document-level assessments were based on sentence-level MT, which does not reflect how the system might perform if fed the entire document as input. To address this, future iterations of the lab could generate two distinct MT outputs: one using sentence-level input (for the first task), and another using the entire document as input (for the document-level task). This would allow students to evaluate how MT quality changes based on input granularity and gain deeper insights into the relationship between system design and translation performance.

The students' reflections not only shed light on the practical challenges and considerations faced during the evaluation process but also highlight the importance of context in determining translation quality. While a few students found it less complex to assess translations at the sentence level, others recognised the added value of considering the broader context provided by document-level evaluation. By encouraging students to critically reflect on their evaluation methodologies and preferences, this exercise served a dual purpose: first, to illustrate how sentence-level evaluations may obscure important context-sensitive errors, and second, to highlight how MT systems often struggle to manage discourse-level phenomena such as reference, coherence, and ambiguity. Through this guided comparison, students became more attuned to the limitations of both evaluation practices and MT outputs, an awareness that is crucial for translators working in increasingly MT-integrated workflows.

 doi:10.20944/preprints202508.0597.v1

In this way, the lab not only supports the development of evaluation literacy but also contributes to broader translation pedagogy by foregrounding the role of context in both human and machine translation assessment.

This exercise provided direct answers to the two research questions posed at the outset. In response to RQ1 (how a context-aware MT evaluation exercise can be effectively implemented in translator training), the findings show that such an exercise can be successfully integrated into a lab setting using a carefully scaffolded structure: beginning with sentence-level evaluation, progressing to document-level assessment, and combining hands-on evaluation tasks with guided reflection. he DELA corpus and spreadsheet-based evaluation tool proved pedagogically effective, and students' continued use of the materials in later coursework suggests sustained engagement.

In response to RQ2 (the impact of context-aware MT evaluation on students' ability to assess translation quality compared to sentence-level evaluation) the results show that the contrast between the two methodologies led students to recognise how sentence-level evaluations can obscure context-dependent issues such as reference, gender, or coherence. Most students expressed increased confidence and satisfaction when evaluating in context, noting that document-level assessment allowed for more accurate judgments and deeper insights into translation quality. This shift in awareness highlights the pedagogical value of exposing students to the contextual blind spots of both MT systems and traditional evaluation practices.

By encouraging students to critically reflect on their evaluation strategies, this exercise not only enhanced their evaluation literacy, but also helped them develop a more nuanced understanding of the complexities involved in assessing MT output. As future practitioners and researchers in the field of translation technology, the insights gained through this comparative methodology will serve them well in evaluating and improving MT systems within diverse professional and academic contexts.

Acknowledgments: This research was funded by the School of Applied Languages and Intercultural Studies at DCU, and partially with the financial support of Science Foundation Ireland at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at DCU [13/RC/2106 P2].

References

1. Kenny, D. Technology and Translator Training. In *The Routledge Handbook of Translation and Technology*; O'Hagan, M., Ed.; Routledge: London and New York, 2019; pp. 498–515.
2. Doherty, S.; Kenny, D.; Way, A. Taking Statistical Machine Translation to the Student Translator. In Proceedings of the Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Commercial MT User Program, 2012.
3. Doherty, S.; Kenny, D. The design and evaluation of a Statistical Machine Translation syllabus for translation students. *The Interpreter and Translator Trainer* **2014**, *8*, 295–315, [https://doi.org/10.1080/1750399X.2014.937571]. https://doi.org/10.1080/1750399X.2014.937571.
4. Mellinger, C.D. Translators and Machine Translation: Knowledge and Skills Gaps in Translator Pedagogy. *The Interpreter and Translator Trainer* **2017**, *11*, 280–293. https://doi.org/10.1080/1750399X.2017.1359760.
5. Castilho, S.; Moorkens, J.; Gaspari, F.; Calixto, I.; Tinsley, J.; Way, A. Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics* **2017**, *108*, 109–120.
6. Tavares, C.; Tallone, L.; Oliveira, L.; Ribeiro, S. The Challenges of Teaching and Assessing Technical Translation in an Era of Neural Machine Translation. *Education Sciences* **2023**, *13*. https://doi.org/10.3390/educsci13060541.
7. Castilho, S.; Knowles, R. A survey of context in neural machine translation and its evaluation. *Natural Language Processing* **2024**, pp. 1–31. https://doi.org/10.1017/nlp.2024.7.
8. Castilho, S.; Doherty, S.; Gaspari, F.; Moorkens, J. Approaches to Human and Machine Translation Quality Assessment. In *Translation Quality Assessment: From Principles to Practice*; Springer International Publishing, 2018; Vol. 1, *Machine Translation: Technologies and Applications*, pp. 9–38.
9. Gaspari, F.; Almaghout, H.; Doherty, S. A Survey of Machine Translation Competences: Insights for Translation Technology Educators and Practitioners. *Perspectives: Studies in Translation Theory and Practice* **2015**, *23*, 333–358. https://doi.org/10.1080/0907676X.2014.979842.

 doi:10.20944/preprints202508.0597.v1

10. Esqueda, M.D. Machine translation: teaching and learning issues. *Trabalhos em Linguística Aplicada* **2021**, *60*, 282–299.
11. Moorkens, J. What to expect from Neural Machine Translation: a practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer* **2018**, *12*, 375–387, [https://doi.org/10.1080/1750399X.2018.1501639]. https://doi.org/10.1080/1750399X.2018.1501639.
12. Guerberof-Arenas, A.; Moorkens, J. Machine Translation and Post-editing Training as Part of a Master's Programme. 2019. Conference paper.
13. Guerberof-Arenas, A.; Valdez, S.; Dorst, A.G. Does Training in Post-editing Affect Creativity? *The Journal of Specialised Translation* **2024**, pp. 74–97.
14. Girletti, S.; Lefer, M.A. Introducing MTPE pricing in translator training: a concrete proposal for MT instructors. *The Interpreter and Translator Trainer* **2024**, *18*, 1–18. https://doi.org/10.1080/1750399X.2023.2299914.
15. Castilho, S.; Mallon, C.Q.; Meister, R.; Yue, S. Do online Machine Translation Systems Care for Context? What About a GPT Model? In Proceedings of the Proceedings of the 24th Annual Conference of the European Association for Machine Translation; Nurminen, M.; Brenner, J.; Koponen, M.; Latomaa, S.; Mikhailov, M.; Schierl, F.; Ranasinghe, T.; Vanmassenhove, E.; Vidal, S.A.; Aranberri, N.; et al., Eds., Tampere, Finland, 2023; pp. 393–417.
16. Läubli, S.; Castilho, S.; Neubig, G.; Sennrich, R.; Shen, Q.; Toral, A. A Set of Recommendations for Assessing Human–Machine Parity in Language Translation. *Journal of Artificial Intelligence Research* **2020**, *67*, 653–672.
17. Kocmi, T.; Avramidis, E.; Bawden, R.; Bojar, O.; Dvorkovich, A.; Federmann, C.; Fishel, M.; Freitag, M.; Gowda, T.; Grundkiewicz, R.; et al. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In Proceedings of the Proceedings of the Eighth Conference on Machine Translation; Koehn, P.; Haddow, B.; Kocmi, T.; Monz, C., Eds., Singapore, 2023; pp. 1–42. https://doi.org/10.18653/v1/2023.wmt-1.1.
18. Castilho, S.; Knowles, R. Editors' foreword. *Natural Language Processing* **2025**, *31*, 983–985. https://doi.org/10.1017/nlp.2024.6.
19. Castilho, S. Towards Document-Level Human MT Evaluation: On the Issues of Annotator Agreement, Effort and Misevaluation. In Proceedings of the Proceedings of the Workshop on Human Evaluation of NLP Systems. Association for Computational Linguistics, 2021, pp. 34–45.
20. Castilho, S.; Popović, M.; Way, A. On Context Span Needed for MT Evaluation. In Proceedings of the Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20), Marseille, France, May 2020; p. 3735–3742.
21. Castilho, S. How Much Context Span is Enough? Examining Context-Related Issues for Document-level MT. In Proceedings of the Proceedings of the Language Resources and Evaluation Conference, Marseille, France, June 2022; pp. 3017–3025.
22. Castilho, S. On the same page? Comparing inter-annotator agreement in sentence and document level human machine translation evaluation **2020**.
23. Freitag, M.; Foster, G.; Grangier, D.; Ratnakar, V.; Tan, Q.; Macherey, W. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics* **2021**, *9*, 1460–1474. https://doi.org/10.1162/tacl_a_00437.
24. Castilho, S.; Knowles, R., Eds. *Special Issue: The Role of Context in Neural Machine Translation Systems and its Evaluation*, Vol. 31, 2025. Special issue editors: Sheila Castilho and Rebecca Knowles.
25. Castilho, S.; Cavalheiro Camargo, J.L.; Menezes, M.; Way, A. DELA Corpus: A Document-Level Corpus Annotated with Context-Related Issues. In Proceedings of the Proceedings of the Sixth Conference on Machine Translation. Association for Computational Linguistics, 2021, pp. 571–582.
26. Lommel, A.; Uszkoreit, H.; Burchardt, A. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica* **2014**, pp. 0455–463.