Perceptions of Educators on MTQA Curriculum and Instruction

João Lucas Cavalheiro Camargo, Sheila Castilho, Joss Moorkens
SALIS/ADAPT Centre
Dublin City University
joao.camargo@adaptcentre.ie
sheila.castilho@dcu.ie, joss.moorkens@dcu.ie

Abstract

This paper reports the results of a survey aimed at identifying and exploring the attitudes and recommendations of machine translation quality assessment (MTQA) educators. Drawing upon elements from the literature on MTQA teaching, the survey explores themes that may pose a challenge or lead to successful implementation of human evaluation, as the literature shows that there has not been enough design and reporting. Results show educators' awareness of the topic, awareness stemming from the recommendations of the literature on MT evaluation, and reports new challenges and issues.

1 Introduction

Academia and industry continuously make efforts to assess the quality of machine translation (MT) systems (Way, 2020), typically using automatic evaluation metrics (AEM) or human evaluation (HE) (Castilho et al., 2018), each approach possessing its own strengths and weaknesses. However, to evaluate an MT system with detailed and actionable results, it is vital to use a balanced approach incorporating HE in the process in conjunction with AEMs (Way, 2020). In particular, the inclusion of HE must be carefully employed so as to not generate hyperbolic reports of the capabilities of MT systems in particular scenarios such as in Hassan et al. (2018).

Some studies have recommended more rigorous HE design principles (Toral et al., 2018; Läubli et al., 2020) not only to dampen hype, but also to identify systems' weaknesses through an analysis of complex linguistic phenomena (Castilho and Caseli, 2023). While it is not recommended to rely solely on AEM-based evaluations (Moorkens, 2022), the literature shows a common tendency to rely on AEMs without HE (Marie et al., 2021; Rivera-Trigueros, 2022) in the MT community. It is understood that MT use must consider the purpose and value of translations and the expected longevity of the content (Way, 2013), which extends to MT evaluation as well (Doherty et al., 2018). In this manner, risks from MT systems such as grammatical errors or inappropriate words/constructions (Koehn and Knowles, 2017), biases in the output (Prates et al., 2020), which can be dangerous for specific domains such as legal and medical (Vieira et al., 2021), can be prevented with rigorous HE incorporation in MT evaluations. Given these risks and the responsibility of implementing a careful evaluation, complementing automatic with HE is essential to ensure AI technology is safe, beneficial and fair (Dignum, 2020). It can be achieved with ethical behaviours adopted by engineers and technology developers (Moorkens, 2022), which can be further refined with the training of stakeholders themselves (Dignum, 2020).

Thus, this paper focuses on the instructional training of MT quality assessment (MTQA), as part of a doctoral study that intends to create and provide training in HE for Natural Language Processing (NLP) master's students. In this paper, we report results from the qualitative findings of a survey aimed at MTQA educators with both TS and NLP educators. It inquired about the educators' attitudes and recommendations regarding HE in MTQA teaching, exploring where HE can be positioned pedagogically, what HE content should

Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), pages 492–506
June 24-27, 2024 ©2024 European Association for Machine Translation

be prioritised, and evaluates the practical considerations that may facilitate or hinder the incorporation of HE into an MTQA curriculum focused on NLP students. The survey explores the following key questions:

1. What are educators' attitudes towards MTQA?
2. What approaches can be taught to foster HE in MTQA?

These findings can inform MTQA trainers and curriculum planners in making informed decisions to foster appropriate HE teaching and deployment, and consequently, its use in MTQA.

2 Related Work

Translation Quality Assessment (TQA) is complex, leading to much debate and different definitions of translation quality, especially in regard to translation technologies, such as MT (Castilho et al., 2018).

MTQA in Translation Studies (TS) curricula has been slowly introduced alongside the use of MT (Korošec, 2011; Dejica-Cartis, 2012) from a curricular standpoint (Doherty and Kenny, 2014) to critically use and assess MT (Rossi, 2017; Moorkens, 2018). Technical aspects of MT also became an element of MTQA teaching, such as building an engine (Farrell and others, 2017), mainly with the intent of empowering trainee translators to understand how the systems work in order to facilitate informed decisions when evaluating the output (Kenny and Doherty, 2014).

Studies have shown that translators in training gain MTQA proficiency through error analysis (Venkatesan, 2018; Loock, 2020), and that translators' ability to identify missing contextual information in MT output and select appropriate language for specific domains is crucial (Núñez, 2019; Bulut, 2019). This mirrors evaluation models used in the industry (Castilho et al., 2018), showcasing academia's efforts to prepare translators.

Accordingly, AEM and other measures of HE have been introduced in the classroom in the TS field. Doherty and Kenny (2014) and Moorkens (2018) introduced adequacy and fluency measures in conjunction with error typologies. Post and Lopez (2014) created a platform on which students could rank MT outputs and generate BLEU scores (Papineni et al., 2002), focusing on the correlation of human judgement with the AEM. Other platforms were used in classroom settings, such as the Asiya-Online toolkit (Giménez and Màrquez, 2010), which provided automatic scores, and later, MutNMT (Ramírez-Sánchez et al., 2021; Ramírez-Sánchez, 2023) for guided building and evaluation of NMT systems. Krüger (2022) proposed Jupyter notebooks to introduce translators to the technical nature of AEMs while generating different scores such as BLEU, METEOR (Banerjee and Lavie, 2005), chrF3 (Popović, 2015), TER (Snover et al., 2006) and BERTScore (Zhang et al., 2019). Macken et al. (2023) demonstrate a case study of teaching MTQA, by using HE through ranking, adequacy and fluency measures, correlating to AEMs provided by MATEO (Vanroy et al., 2023), a platform that generates BLEU, ChrF, BERTScore, BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020) scores. These studies demonstrate the effort to introduce different evaluation approaches in the translation classroom, and how important accessible evaluation platforms are for training translators.

The importance of integrating MTQA into curricula is underscored by the concept of MT Literacy (Bowker and Ciro, 2019) which entails understanding the operational mechanisms of MT systems to facilitate their use. Krüger (2022) and Macken et al. (2023) echo the importance of MT literacy in equipping professionals to use and evaluate MT effectively. However, the implementation of training is context-based, the pedagogical guiding principles for MTQA education tend not to be structured.

In the context of NLP education, the few studies that mention MTQA do so only to a minor degree (Alm et al., 2016; Martynova et al., 2018; Artemova et al., 2021). This is due to MT being only one component within the broader spectrum of training, with evaluation assuming a secondary role. However, that does not diminish the importance of evaluation in NLP, as the reasons for its lack of implementation in training may due to absence of space in the curriculum and the lack of structured information on evaluation (Madureira, 2021). As such, organising the insights and recommendations of MTQA educators, both from NLP and TS may lead to fostering MTQA education.

3 Methods

To collect information regarding educators' insights and suggestions on MTQA, an online sur-

vey was designed (approved by the university's Research Ethics Committee, with reference DCU-FHSS-2023-015).

3.1 Design

The survey was created on the platform Qualtrics.[1] It was formulated with closed-ended and open-ended questions, divided in blocks:[2]

- the plain language statement and consent form[3]
- 13 questions related to the profile of the educators
- four questions related to opinions and attitudes regarding HE
- 11 questions related to general MTQA pedagogy
- six questions related to recommendations of HE for MTQA

3.2 Participants

The participants are MTQA educators from different fields, such as TS, NLP and Computational Linguistics (CL).[4]

The participants were recruited via: i) social media; ii) email via a curriculum analysis examining universities' postgraduate programmes in Europe and; iii) email collection by examining publications related to MTQA teaching. Note that participants data was anonymised.

4 Data Results and Analysis

As this is an ongoing study, the results reported in this paper are qualitative and small-scale, with the intention of being exploratory, to explore possible relationships and patterns (Cohen et al., 2017). While it is known that smaller samples are not ideal for generalisations (Saldanha and O'Brien, 2014), the qualitative components may inform better the results of the survey as it reaches a larger-scale (McMillan and Schumacher, 2010).

Data was visualised on Qualtrics, which affords analysis of both closed-ended and open-ended questions. For the closed-ended questions, Qualtrics automatically created graphs based on the responses to form variables, and the platform allowed a degree of customisation to change the colour of graphs and combine/separate variables (or groups) as needed. For the open-ended questions, Qualtrics lists the responses by variables (or groups), allowing an interpretive qualitative analysis of the data.

4.1 Participants' Background

Data drawn from 27 participants were analysed.

Q1 - What is your field? Participants could choose multiple fields to accommodate interdisciplinarity among the educators. 18 participants chose 'Translation Studies' as their field of teaching, five participants chose 'Computational Linguistics', seven participants chose 'Natural Language Processing'.

Among the 28 participants, one participant added 'Speech Processing' as their field, one participant added Human-Computer Interaction as their field and another added 'Computer Science 'via the 'other' option. While CL and NLP may have often been used interchangeably in research, they represent different streams of research with different emphases, as Tsujii (2011) demonstrates with their experiment. We also acknowledge that the boundaries may not easily be defined (Luz, 2022). Therefore, methodologically we make no distinction between these two groups, and to aid visualisation, the responses from NLP/CL will be organised and reported as a single group, as such, this leads to nine participants in the NLP/CL group.

4.2 Types of MTQA

Participants were asked about the type of evaluation they teach by answering the question:

Q2 - What types of MT evaluation do you teach? As can be seen in Figure 1, the TS group mostly teaches HE, followed by AEMs and semi-automatic evaluation. When prompted in a follow-up question to explain their comments, the TS educators explained their experience:

- One participant notes that MT evaluation is taught to foster MT literacy leading to better use of the systems.
- One participant has PE as the central type of evaluation, while also teaching HE and AEM to a lesser extent.
- One participant focuses on evaluation through PE.

[1] Available on: `https://www.qualtrics.com`

[2] The full questionnaire can be found in Appendix A.

[3] This explained the research aims, the ethical aspects and how the data is handled

[4] The distinctions between CL and NLP was made to accommodate possible different curricular nomenclature and personal preferences.

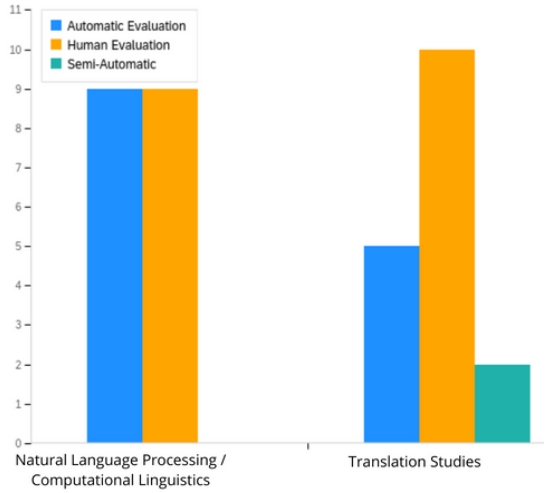

Figure 1: Types of evaluation TS and NLP/CL educators teach

- One participant considers HE the focus of the lesson using the DQF-MQM framework.
- One participant distinguishes MQM from HE, where focus is on MQM, but also mentioning other HE methods, and minor emphasis to AEMs.

Within the NLP/CL group, there are equal efforts reported into teaching HE and AEMs. Four participants described more about their teaching:

- One participant mention teaching HE and AEMs (BLEU, BERT and Comet).
- One participant mentions MTQA is only a component of the course.
- One participant teaches different metrics to different groups. For their Master's students in Artificial Intelligence, they teach AEMs. For undergraduate translation students, they teach HE.
- One participant mentions teaching AEMs very briefly to make students understand their use in the context of testing the development of a system.

4.3 Attitudes Towards Human Evaluation

This subsection explores participants' expectations and attitudes towards HE (Q3 and Q4)

Q3 - In your opinion, what trends do you foresee in evaluation metrics that incorporate human judgment for MT systems? Select all that apply. As may be seen in Figure 2 for the TS group, the most commonly-selected options were context for Quality Assessment (QA), customised evaluation, an equal amount for User Experience (UX) evaluation and multimodal approaches, followed by ethics, crowdsourced evaluation and two 'Other' responses. These two responses were 'comparing several systems with emphasis on output' and another response said that all the topics could be important except for crowdsourcing. From NLP/CL, the most commonly-selected were ethics and customised evaluation. Followed by an equal selection of UX evaluation and context-based evaluation. It is worth noting that crowdsourced evaluation was not chosen among the NLP/CL group, which is surprising as the field is known to use crowdworkers for evaluation. The bigger focus given to ethics supports Moorkens' (2022) assertion that bigger emphasis must be given to the ethical behaviours of engineers, possibly showing that NLP/CL teachers are aware of this. One participant chose 'Other' to suggest the use of Large Language Models (LLM) to emulate HE.

Q4 - In your view, what constitutes a comprehensive evaluation of an MT system? Please describe the key components or criteria that should be included. From TS, nine responses focused only on human judgements and six responses included the use of AEM combined with HE. From NLP/CL, six responses mentioned only human judgements, two responses mentioned a combination of AEM with HE and among the eight answers, four mentioned evaluating MT systems for a specific purpose.[5] The responses from the TS group mentioned:

- Combined measures of HE and AEM, with State Of The Art (SOTA) metrics, and their correlation.
- Evaluation with platforms with good UX (clean interface, resembling the working environment of a translator).
- Genre, style, terminology, purpose of the text, and agreement with the clients' needs.
- Use of DQF-MQM for measuring error typology.
- Different degrees of use of MT output, from raw MT to PE at different levels.
- Evaluations that consider human translations as references.
- Measurement of technical aspects (such as training data, speed, pricing, pollution).

The attitudes from the TS group echoes some of the expectations from the industry, such as the adoption of TQA frameworks such as DQF-MQM,

[5]The full qualitative results are included in Appendix B.

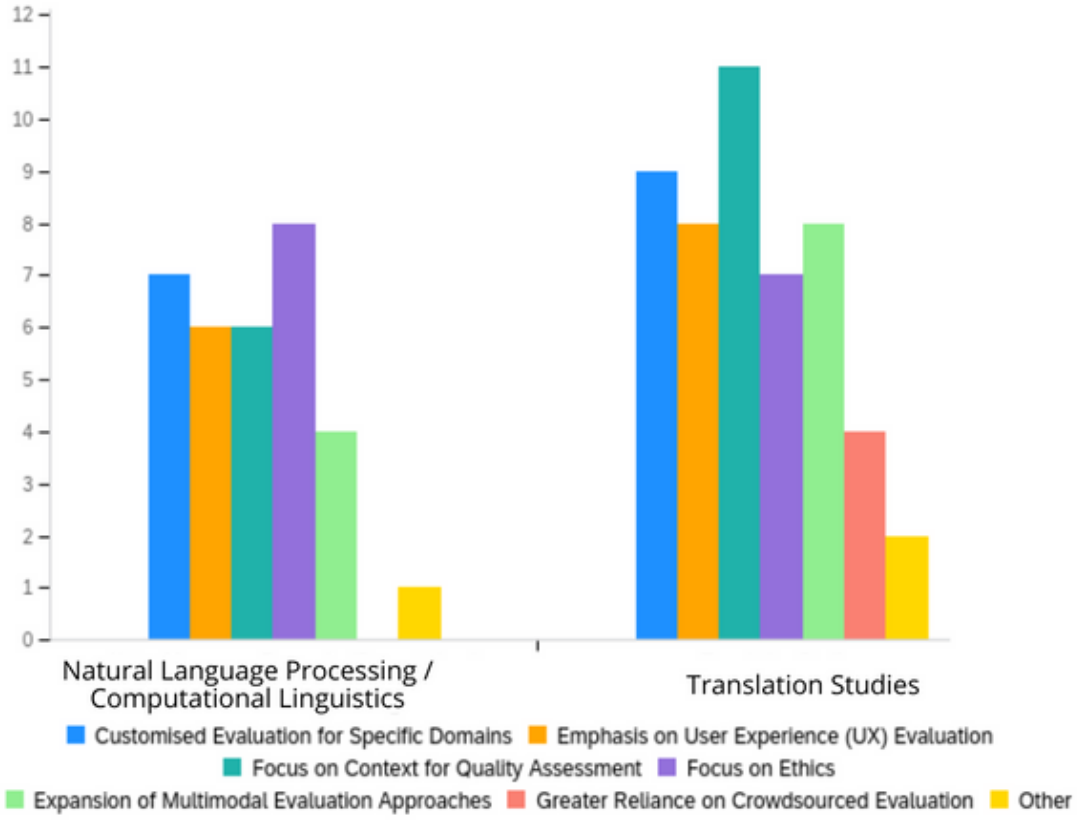

Figure 2: Future trends of Evaluation chosen by TS and NLP/CL teachers

pricing and productivity expectations in addition to the expectation of clients' needs (Castilho et al., 2018). While the NLP/CL group cites the following:

- HE measures such as adequacy, fluency, error analysis and different classifications
- Document-level considerations, such as cohesion and coherence.
- A combination of HE and AEMs, but ultimately with task-based evaluation in mind, to consider how good is the MT system for its appropriate use.
- Risk assessment, considering the type of errors and their severity, according to the domain.
- User-centred assessment, where the end user's purpose of using the translation is to complete a task or is satisfied by its use.

The perishability of content and its purpose (Way, 2013; Way, 2020), in addition to risk assessment which should increasingly be introduced in the training (Doherty et al., 2018) can be noticed by the results of these expected trends. Further, document-level considerations also follow the recommendations made for MT evaluation (Läubli et al., 2020).

4.4 Pedagogical Factors and Recommendations for MTQA

This subsection focuses on the central aspect of MTQA teaching and NLP education (Q5, Q6, and Q7).

Q5 - Assess the importance of including Evaluation Metrics in your academic curriculum - In response to this question, participants assessed the inclusion of both AEM and HE in their teaching curriculum, as can be seen in Figure 3.

Regarding AEMs, the consensus among the NLP/CL participants were that AEMs are *'extremely important'*, while for TS the most chosen option was *'moderately important'*. Regarding HE, while all groups claimed it to be *'extremely important"*, the TS group mentioned that the emphasis is on HE since they are teaching translators, and therefore AEMs are given less focus. The NLP/CL group mentioned the importance of both AEMs and HE. Interestingly one participant of the TS group mentioned that AEMs are equally important, and one NLP/CL participant stated that, since the course they teach is technical, less emphasis is given to HE.

Following Q5, participants were able to add

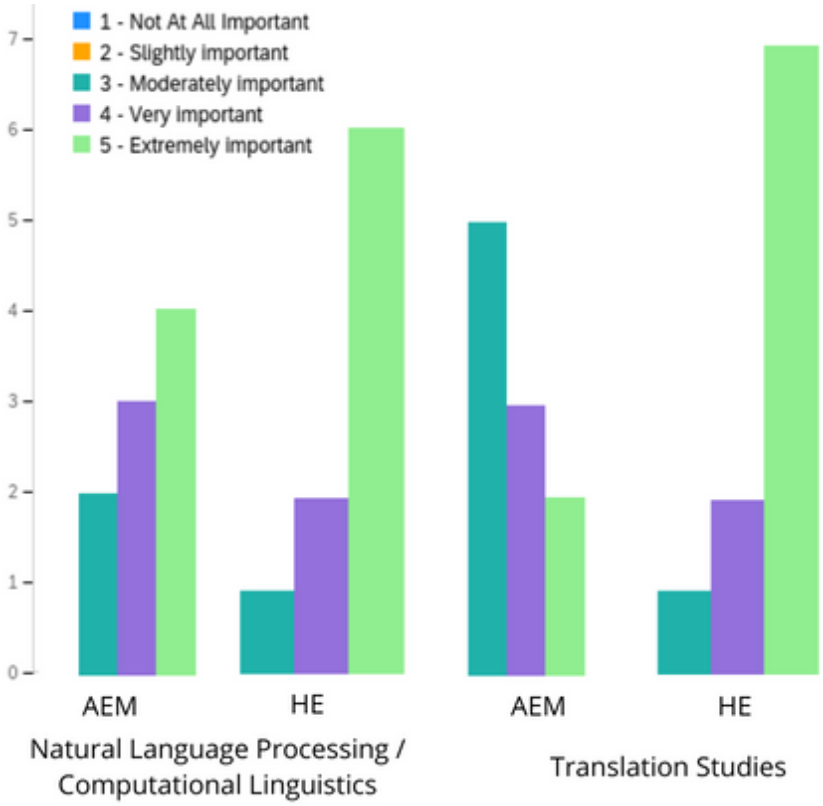

Figure 3: Importance of including AEMs and HE in the curriculum responded by MTQA teachers

comments by responding to ***Q6 - Please, add any further comments or explanations for your previous answer***. In the TS group, a participant explained that contextually it is more valuable for them to teach HE towards translators, as AEMs are given less focus. Another participant emphasised that the type of student and level matters when teaching each type of metric. For such participant, undergraduate students who are studying to be come translators may require less attention towards both metrics, but the educator explains that for master's NLP students there is room to introduce it to them.

In the NLP/CL group, two participants explained that both metrics are relevant, AEMs providing fast, cheap and objective system checks, while HE are used to understand the values of AEMs and providing insights to improve the systems. One participant differentiates the teaching of metrics in two ways: the first, being moderately important, teaching the metrics directly (such as adequacy scores, error annotation for HE and AEMs such as COMET); while another participant mentions that the most important is to teach the general concepts of HE and AEMs in detail - alluding to a better understanding of the evaluation process as more important than teaching individual metrics. One participant comments that considering they teach more technical courses, there is less focus on HE.

Both groups correspond to the expectations to a curriculum focused on MT and its evaluation, as what matters the most is the context in which they are inserted (Kenny and Doherty, 2014), whether they are translators or developers, but not forgetting th

Q7 - If you were to create a Human Evaluation module in MT quality assessment addressed to NLP students at Master's level, what should be the main content? Select as many as necessary. In this question participants gave their opinion on the important contents to be taught, as seen in Figure 4.

From the group TS, the most widely chosen option was translators as expert evaluators and design of MT evaluation, followed by adequacy/fluency measures and error typology. The responses from TS may follow the recommendations from the literature such as Laubli et al. (2020) and overall correspond to the importance given to translators (Kenny and Doherty, 2014), such as advisors on the evaluation process (Moorkens, 2017).

From the NLP/CL group, the most widely chosen options were adequacy/fluency measures and inter-annotator agreement, followed by usability and design of MT evaluation. When asked to add other topics (if any), one participant from TS suggested understanding evaluation tools and platforms with analytics, and another TS educator suggested how to interpret results, including generalisability. Among the NLP/CL group, one participant suggested that a whole module on evaluation is not justified, and another participant suggested 'mid-level evaluators', reproducibility of evaluation and bias detection.

Further recommendations in the realm of UX are given, as one of the participants mention how tools and platforms with analytics and insights are important to be introduced, especially if they are accessible. This may be a reason why platforms such as MATEO are being adopted in the classroom (Macken et al., 2023), and to avoid issues that had happened before as reported in Doherty and Kenny (2014) when students were not able to perform AEM scoring due to the unfriendliness of the platforms.

4.5 Pedagogical Challenges

The literature indicates different reasons that may impede more training on evaluation, such as the curriculum (Madureira, 2021) or limited motivation to perform and understand QA processes (Doherty et al., 2018). Thus, this section focuses on the pedagogical elements that may introduce problems in implementing MTQA teaching.

Q8 - Beyond content (such as human evaluation metrics or automatic evaluation metrics), what other pedagogical aspects do you believe may be currently lacking in the teaching of MT quality assessment? Please select all that apply. Participants could select different aspects of teaching such as instructional constraints, hours, and others, as seen in Figure 5

In the TS group, the most commonly-chosen options were allocated hours and faculty expertise and development, followed by curricular structure and lastly by scalability of teaching methods. Expertise and development being one of the most chosen resonates with Doherty et al. (2018) mentioning how educators have to face an evolving and rapidly changing technological scenario, which may make teaching MTQA more difficult. The allocated hours being also one of the most chosen might be related to MTQA being taught under modules on translation technologies where MT is one component and MTQA is a minor aspect, or a module focused on MT which covers different paradigms, use-cases and MTQA may have more room.

In the NLP/CL group, the most commonly-chosen was allocated hours followed by curricular structure followed by the allocated hours, which has been seen in the literature beforehand as an issue (Madureira, 2021).

Q9 - If you were to create a Human Evaluation module in MT quality assessment addressed to NLP students at Master's level, what would be the best format? Inquired about an ideal format for MTQA training focused on HE, participants responded the following as per Figure

The TS group by a majority suggested an academic module (which is probably unlikely given the previously mentioned time constraints within programmes), followed by the option of a week-long course and a workshop. The NLP/CL group suggested equally an academic module and the option 'other', followed by a two-day course and a week-long course. The 'other' response suggested that each format could be taught depending on the purpose, such as the massive open online course in order to have more time, or a whole-day workshop to introduce the basis of evaluation, or in between a two-day and a week long course, leading the learning to be more contextual. As a follow up, they were asked a question about modality.

Q10 - Given your previous choice on the best format for a Human Evaluation module, what teaching modality would be most suitable? As seen in figure 7, the TS group chose in-person, spread out over several weeks, followed by blended, with the least chosen as an online, synchronous training. Most of the NLP/CL group chose an in-person intensive training, followed by and online synchronous training and an in-person training spread out over several weeks. The 'other' option chosen by a participant of the NLP/CL group suggested that the best modality depends more on the teacher than the topic itself.

Q11 - Please, add any further comments or explanations for your previous answers from Q10 and Q09 here. Within the TS group, one participant commented that the in-person contact is important for the possibility of providing technical

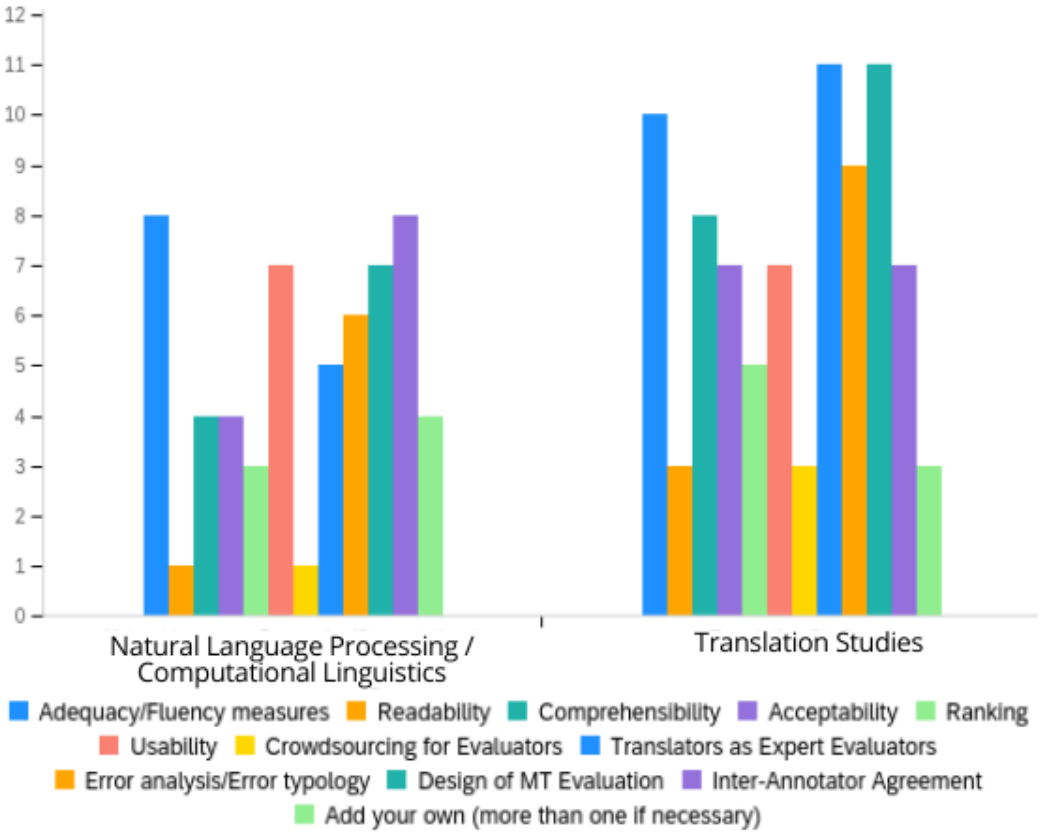

Figure 4: Human Evaluation methods and metrics divided among TS and NLP

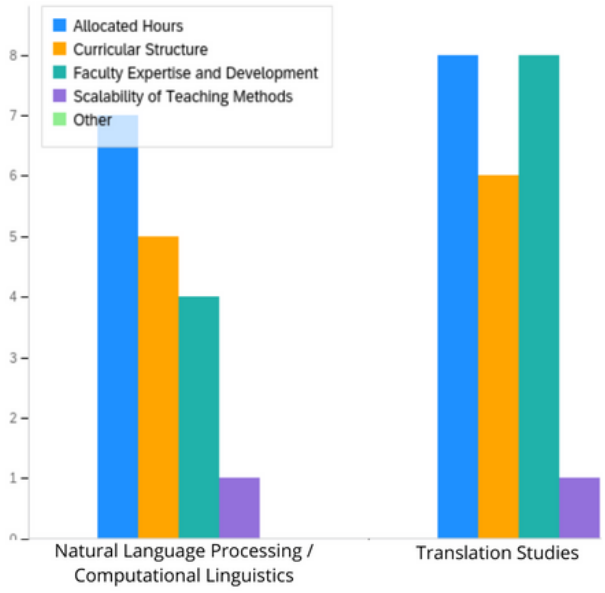

Figure 5: Pedagogical constraints in MTQA teaching divided by TS and NLP/CL

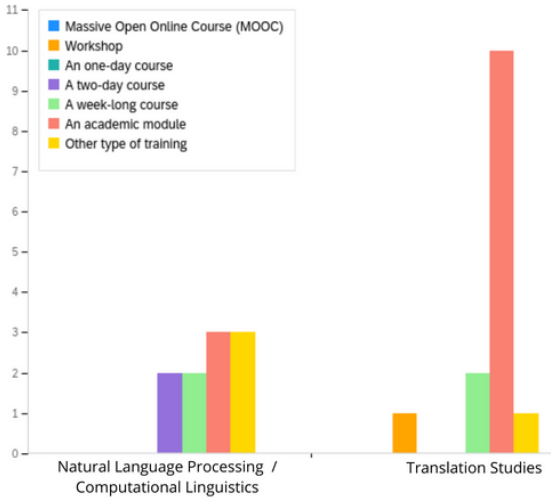

Figure 6: Preferred format for HE in MTQA training responded by TS and NLP/CL educators

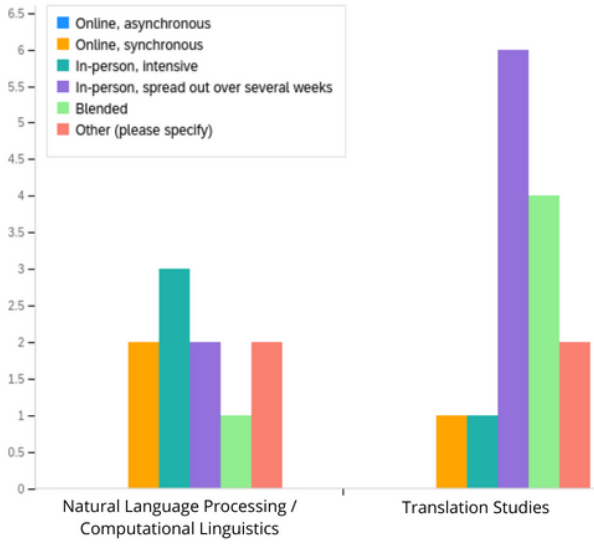

Figure 7: Preferred modality for HE training responded by TS and NLP/CL teachers

support, alluding to easier technical support to students with certain aspects of MTQA. Another participant explained in detail about their experience for a Master's level training, addressing that academic modules are the only mandatory elements, so the participant suggests spread-out hands-on sessions, such as workshops, in order to provide the different aspects of evaluation for NLP/CL students. Another TS educator complemented that since translator competence takes time to develop, translation evaluation also follows, thus, advocating for long-term training. One educator emphasises that understanding and agreement with the needs of the students would be important to choose the format, so long there was interaction. While the NLP/CL group suggests as many laboratory and hands-on sessions as possible, while another educator suggests that long-term training spanning overall several weeks allow discussion and the opportunity of individual work.

5 Final Considerations and Future Challenges

By comparing the two groups, it can be seen that their attitudes and difficulties reflect both contextual factors of their teaching, and needs commonly associated with their profession.

For TS educators, there has been an increasing effort to integrate the newest technological advancements into their teaching while still maintaining the critical approach of their use. TS educators focus on teaching MTQA for translation trainees in order to foster their MT literacy, either for more proficient use when performing PE or to prepare them to serve as consultants in the development of MT systems. For either, it places TS educators and the future translators in a position to ensure a safer use of translation systems. NLP/CL educators tend to place more attention towards the ethics, and regard the design of MT evaluation among the most chosen topics, which can be performed by translators who can serve as experts on this process.

We have seen in section 4.5 that the technical aspect may present different pedagogical challenges for TS educators **Q8**, since teaching technical elements to a non-technical audience requires accessible resources. Therefore, there has been research done focused on the experience of translators performing PE, and evaluating MT systems. As a result, over the past years platforms such as MutNMT and MATEO are paramount to make aspects of evaluation accessible, especially when teaching AEMs. Accordingly, for TS lecturers who reported faculty expertise and development as a pedagogical difficulty, those platforms are an important resource for educators.

The NLP/CL group reports other difficulties with MTQA, primarily in finding room in the curricular structure to focus on evaluation (Section (4.5, **Q8**). It is worth noting that this group does not recognise either type of evaluation as less important. In fact, the survey shows that NLP/CL educators recognise the importance of HE in MTQA and teach different evaluation metrics to different groups according to their profiles and roles in the evaluation process. Due to the amount of technical content in development to be covered, it has been suggested by NLP/CL educators that the most appropriate way to cover evaluation would be through intensive, interactive, hands-on workshops to practise different aspects of evaluation - either the design planning, different approaches or the annotation. However, based on the results, NLP/CL educators appear to suggest that NLP master's students who are choosing to work on MT development and evaluation should know the basic approaches and should still place translators at the centre of the evaluation. These results show the efforts of the MT community at demonstrating the importance of every stakeholder in the MTQA process - from developer to evaluator.

The design and implementation of MTQA still

brings challenges (Section 4.5), but TS, NLP and CL educators report it is essential, whether you are training translators or developers. Challenges to overcome may include:

- Teaching the design of a MT system evaluation is important, but also the user-friendliness of the platform or methodology of evaluators, placing UX as a worthy topic to investigate.
- Finding space in the curriculum for evaluation may be difficult, so a solution proposed is the design, development and implementation of practical workshops around MTQA.
- LLM-based evaluations emulating HE may become more common, and thus, educators need to be prepared to teach NLP professionals the appropriateness of using this approach in evaluation.

As observed in section 4.3, (**Q4**), the survey also provided some insights on what constitutes a comprehensive evaluation of MT, demonstrating the awareness of the educators.

- Due to its situational nature, the purpose of the system and its end-user are important factors in designing an evaluation.
- A combination of HE and AEM and its correlations are ideal, particularly to show in training.
- Risk assessment and perishability of content are a factor to note the degree of how comprehensive the evaluation should be.

This survey shines light on the directions of MTQA education according to educators from different fields. We hope the insights and recommendations presented here can aid the MT community in fostering MTQA education.

Acknowledgements: We would like to thank the participants for participating in this research. It is a voluntary survey and the results have shown how busy educators are, we present the utmost gratitude and hope the results are able to provide useful pedagogical insights.

Funding: This research was funded by the School of Applied Language and Intercultural Studies at Dublin City University and with the financial support of Science Foundation Ireland at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at Dublin City University [grant number 13/RC/2106_P2].

References

Alm, Cecilia Ovesdotter, Kathryn Womack, Anne Haake, and Timothy Engström. 2016. A pedagogical model for computational linguistics across curricular boundaries. *Language and Linguistics Compass*, 10(7):335–345.

Artemova, Ekaterina, Murat Apishev, Veronika Sarkisyan, Sergey Aksenov, Denis Kirjanov, and Oleg Serikov. 2021. Teaching a massive open online course on natural language processing. *arXiv preprint arXiv:2104.12846*.

Banerjee, Satanjeev and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Bowker, Lynne and Jairo Buitrago Ciro. 2019. *Machine translation and global research: Towards improved machine translation literacy in the scholarly community*. Emerald Publishing Limited.

Bulut, Senem ÖNER. 2019. Integrating machine translation into translator training: towards 'human translator competence'? *transLogos Translation Studies Journal*, 2(2):1–26.

Castilho, Sheila. and Helena de Medeiros Caseli Caseli. 2023. Tradução automática. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*.

Castilho, Sheila, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. Approaches to human and machine translation quality assessment. *Translation quality assessment: From principles to practice*, pages 9–38.

Cohen, Louis, Lawrence Manion, and Keith Morrison. 2017. *Research Methods in Education*. Routledge.

Dejica-Cartis, Daniel. 2012. Developing the electronic tools for translators syllabus at politehnica university of timisoara. *Procedia-Social and Behavioral Sciences*, 46:3614–3618.

Dignum, Virginia. 2020. Responsibility and artificial intelligence. *The oxford handbook of ethics of AI*, 4698:215.

Doherty, Stephen and Dorothy Kenny. 2014. The design and evaluation of a statistical machine translation syllabus for translation students. *The Interpreter and Translator Trainer*, 8(2):295–315.

Doherty, Stephen, Joss Moorkens, Federico Gaspari, and Sheila Castilho. 2018. On education and training in translation quality assessment. *Translation quality assessment: From principles to practice*, pages 95–106.

Farrell, Michael et al. 2017. Building a custom machine translation engine as part of a postgraduate university course: a case study. In *Proceedings of the 39th Conference Translating and the Computer*, pages 35–39.

Giménez, Jesús and Lluís Màrquez. 2010. Asiya: An open toolkit for automatic machine translation (meta-) evaluation. *Fifth Machine Translation Marathon*, 94.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Kenny, Dorothy and Stephen Doherty. 2014. Statistical machine translation in the translation curriculum: overcoming obstacles and empowering translators. *The Interpreter and translator trainer*, 8(2):276–294.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Korošec, Melita Koletnik. 2011. Applicability and challenges of using machine translation in translator training. *ELOPE: English Language Overseas Perspectives and Enquiries*, 8(2):7–18.

Krüger, Ralph. 2022. Using jupyter notebooks as didactic instruments in translation technology teaching. *The Interpreter and Translator Trainer*, 16(4):503–523.

Läubli, Samuel, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of artificial intelligence research*, 67:653–672.

Loock, Rudy. 2020. No more rage against the machine: how the corpus-based identification of machine-translationese can lead to student empowerment. *The Journal of specialised translation (JoSTrans)*, 34:150–170.

Luz, Saturnino. 2022. Computational linguistics and natural language processing. *The Routledge Handbook of Translation and Methodology*, pages 373–391.

Macken, Lieve, Bram Vanroy, and Arda Tezcan. 2023. Adapting machine translation education to the neural era: A case study of mt quality assessment. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 305–314.

Madureira, Brielen. 2021. Flamingos and hedgehogs in the croquet-ground: Teaching evaluation of nlp systems for undergraduate students. In *Proceedings of the Fifth Workshop on Teaching NLP*, pages 87–91.

Marie, Benjamin, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. *arXiv preprint arXiv:2106.15195*, 1(1):7297–7306.

Martynova, Irina, Lilia Metelkova, Natalia Gordeeva, Larisa Nikitinskaya, Margarita Emelianova, and Alena Trukova. 2018. The programs of computational linguistics graduate in german and us universities. *Visnyk Natsional'noi akademii kerivnykh kadriv kultury i mystetstv*, 1(3).

McMillan, James H and Sally Schumacher. 2010. *Research in education: Evidence-based inquiry*. pearson.

Moorkens, Joss. 2017. Under pressure: translation in times of austerity. *Perspectives*, 25(3):464–477.

Moorkens, Joss. 2018. What to expect from neural machine translation: a practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer*, 12(4):375–387.

Moorkens, Joss. 2022. Ethics and machine translation. *Machine translation for everyone*, page 121.

Núñez, Kenneth Jordan. 2019. Análisis de la percepción, la utilidad y la calidad de los sistemas de ta por parte del traductor en formación. *E-Aesla*, 1(5):391–399.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Popović, Maja. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Post, Matt and Adam Lopez. 2014. The machine translation leaderboard. *Prague Bull. Math. Linguistics*, 102:37–46.

Prates, Marcelo OR, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381.

Ramírez-Sánchez, Gema, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Caroline Rossi, Dorothy Kenny, Riccardo Superbo, Pilar Sánchez-Gijón, and Olga Torres-Hostench. 2021. Multitrainmt: training materials to approach neural machine translation from scratch. In *TRITON 2021 (Translation and Interpreting Technology Online)*.

Ramírez-Sánchez, Gema. 2023. Mutnmt, an open-source nmt tool for educational purposes. In *Proceedings of the 1st Workshop on Open Community-Driven Machine Translation*.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Rivera-Trigueros, Irene. 2022. Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, 56(2):593–619.

Rossi, Caroline. 2017. Introducing statistical machine translation in translator training: from uses and perceptions to course design, and back again. *Revista Tradumàtica: tecnologies de la traducció*, 1(15):48.

Saldanha, Gabriela and Sharon O'Brien. 2014. *Research methodologies in translation studies*. Routledge.

Sellam, Thibault, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432*.

Tsujii, Jun'ichi. 2011. Computational linguistics and natural language processing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 52–67. Springer.

Vanroy, Bram, Arda Tezcan, and Lieve Macken. 2023. Mateo: Machine translation evaluation online. In *The 24th Annual Conference of The European Association for Machine Translation (EAMT 2023)*, pages 499–500. European Association for Machine Translation (EAMT).

Venkatesan, Hari. 2018. Teaching translation in the age of neural machine translation. *APLX 2017 at Taipei Tech-Transformation and Development: Language, Culture, Pedagogy and Translation*, pages 39–54.

Vieira, Lucas Nunes, Minako O'Hagan, and Carol O'Sullivan. 2021. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11):1515–1532.

Way, Andy. 2013. Emerging use-cases for machine translation. In *Proceedings of Translating and the Computer 35*.

Way, Andy. 2020. Machine translation: Where are we at today. *The Bloomsbury companion to language industry studies*, 1(1):311–332.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix A. Full Questionnaire

Questions in bold are the ones selected for this paper.

- Q1 - Name - (Open-ended)
- Q2 - Email - (Open-ended)
- Q3 - List of Countries - (Close-ended)
- Q4 - What is your highest level of education? - (Close-ended)
- **Q5 (In the survey, Q1) - What is your field?** - (Close-ended)
- Q6 - How many hours do you spend teaching per week? Move the slider according to the amount of hours. - (Close-ended)
- Q7 - What are your other main work activities? - (Close-ended)
- Q8 - In your current teaching role, how much influence do you have over the curriculum, including changes to the syllabus and teaching methods? - (Close-ended)
- Q9 - Please, add any further comments or explanations for your previous answer here. - (Open-ended)
- Q10 - What type of students do you work with, mostly? - (Close-ended)
- Q11 - At what academic levels do you currently teach? Please select all that apply. - (Close-ended)
- Q12 - What modality/modalities do you deliver training in? - (Close-ended)
- Q13 - Have you taught MT quality assessment before? - (Close-ended)
- Q14 - Please rate the significance of incorporating human evaluation into the development of MT systems. Rate on a scale of 1 to 5. - (Close-ended)
- Q15 - Please, add any further comments or explanations for your previous answer here. - (Open-ended)
- **Q16 (In the survey, Q3) - In your opinion, what trends do you foresee in evaluation metrics that incorporate human judgment for MT systems? Select all that apply.** - (Close-ended)
- **Q17 (In the survey, Q4) - In your view, what constitutes a comprehensive evaluation of an MT system? Please describe the key components or criteria that should be included.** - (Open-ended)
- **Q18 (In the survey, Q2) - What types of MT evaluation do you teach?** - (Close-ended)
- Q19 - Please, add any further comments or explanations for your previous answer here. - (Open-ended)
- Q20 - Do you teach evaluation for NLP tasks (e.g. summarisation, speech recognition, sentiment analysis) other than MT? - (Close-ended)
- Q21 - Please, add any further comments or explanations for your previous answer here. - (Open-ended)
- Q22 - How many years have you been teaching MT quality assessment? Move the slider according to the amount of years. - (Close-ended)
- Q23 - Assess the importance of teaching students how to plan evaluations for MT systems in your academic curriculum. Please rate the importance of integrating evaluation planning as part of the academic curriculum for MT quality assessment. - (Close-ended)
- Q24 - Please, add any further comments or explanations for your previous answer here. - (Open-ended)
- **Q25 (In the survey, Q5) - Assess the importance of including Evaluation Metrics in your academic curriculum. Please evaluate the importance of integrating evaluation metrics into the academic curriculum for MT quality assessment. You will be presented with two types of evaluation metrics. Rate on a scale of 1 to 5.** - (Close-ended)
- **Q26 (In the survey, Q6) - Please, add any further comments or explanations for your previous answer here.** - (Open-ended)
- **Q27 (In the survey, Q8) - Beyond content (such as human evaluation metrics or automatic evaluation metrics), what other ped-**

agogical aspects do you believe may be currently lacking in the teaching of MT quality assessment? Please select all that apply. - (Close-ended)

- Q28 - Please, add any further comments or explanations for your previous answer here. - (Open-ended)
- **Q29 (In the survey, Q7) - If you were to create a Human Evaluation module in MT quality assessment addressed to NLP students at Master's level, what should be the main content? Select as many as necessary.** - (Close-ended)
- Q30 - Please, add any further comments or explanations for your previous answer. - (Open-ended)
- **Q31 (In the survey, Q9) - If you were to create a Human Evaluation module in MT quality assessment addressed to NLP students at Master's level, what would be the best format? - (Close-ended)**
- **Q32 (In the survey, Q10) - Given your previous choice on the best format for a Human Evaluation module, what teaching modality would be most suitable? - (Close-ended)**
- **Q33 (In the survey, Q11, in relation to Q9 and Q10) - Please, add any further comments or explanations for your previous answers from Q31 and Q32 here.** - (Open-ended)
- Q34 - Is there anything else you would like to add? - (Open-ended)

Appendix B. Full responses from Q4 - In your view, what constitutes a comprehensive evaluation of an MT system? Please describe the key components or criteria that should be included.

- P1 - adequacy, error annotation and some classification
- P2 - To evaluate an MT system, we should take into account the training data used (quantity and quality) - this includes the pretraining data if the model is based on a pretrained model-, the size of the model (number of parameters), the memory footprint, the speed (inference time). The generalization power and particularly the robustness to domain shift should be evaluated.
- P3 - I think that the evaluation of an MT system cannot be detached from the intended purpose. If the MT system is used to generate draft translations the key thing to evaluate is translation productivity. In the MT system is used for gisting, the key thing to evaluate is the ability of the user of the MT system to perform a task after reading the MT output.
- P4 - Human and automatic evaluation. But ultimately, task-based evaluation is most important: how good is the MT for whom in what situation?
- P5 - For assessing the appropriateness of an MT system, I consider that there are different elements worth considering: 1. The domain of use (e.g. medical, legal, etc.) 2. Translation quality (does the MT system provide "good enough" quality for the domain?) 3. The machine translation user experience (MTUX) (Is a translator the one using the MT system? Any other type of MT user? What are the MT needs of this type of user? Undoubtedly, MT needs will vary among different MT users) Once all these elements have been considered and factored in, an informed decision can be taken, whether X system is appropriate or not for a specific use-case and user type
- P6 - Error analysis, Style preservation, Coherence, Document level aspects
- P7 - A comprehensive evaluation of the usefulness (sometimes called "quality") of an MT system should mimic as much as possible the usage scenario and the indicators of usefulness. For instance, if one wants to use MT to increase the productivity of translators, then evaluation should measure productivity in a scenario which is as similar as possible to that in which translators work. Judging "translation quality" through human judgements (usually produced "in vacuo") is clearly inferior to this approach.
- P8 - 'traditional' sentence-level assessment - document-level assessment - user-centered

assessment : does the translation enable readers to complete a task or otherwise 'satisfies' readers? - error analysis: what type of errors we see, what severity they present, and consequently perform a risk assessment, depending on the type of document and the type of errors found

- P9 - * Oriented to particular MT use (assimilation or dissemination, for example) * Blinded in the sense that humans do not know whether they are evaluating other humans or machines to avoid biases * Measuring productivity in case of MT used by professional translators
- P10 - source text as well as output evaluation
- P11 - Accuracy and style
- P12 - Translation quality assesment, i.e. MT vs human output; evaluation of PE effort; consistent teminology, style; error typology (and several other aspects that I am unaware of at this time and/or may arise in the future)
- P13 - The evaluation should take into account accuracy, appropriateness (genre, stlye, terminology, etc.), general language quality, alignment with clients' needs.
- P14 - Language level. Choice of terminology. Expression of ideolect. Stylistic clarity. Degree of understanding of the sociolect of the translation. Y
- P15 - I think both automatic evaluation and human evaluation are essential. Automatic evaluation should be performed with a sufficiently large sample using one or more SOTA metrics. Human evaluation should be performed in a platform that facilitate scoring with a clean interface and should mimic as much as possible the working environment of a translator.
- P16 - For an evaluation to be comprehensive, it should cover the multiple dimensions involved in the adequacy of the system, from technical aspects (training data, speed, pricing, pollution...) and linguistic (accuracy, fluency, grammaticality, contextual adequacy...) to the user experience (perception, use, ethics...).
- P17 - Evaluation based on both automatic scores and human judgement, as well as investigations into how well they correlate. Comprehensive human evaluation should include error annotation using an error typology such as MQM, ranking tasks and post-editing.
- P18 - Combination of state-of-the-art automatic metrics and human evaluation, including inter-annotator agreement.
- P19 - Beyond the above (usability, context, ethics, multimodality): adequacy metrics, quality-level differentiation, workflow integrability, data transparency
- P20 - Actually, DQF-MQM is a good example of a comprehensive evaluation of MT system.
- P21 - Accuracy and fluency are basic metrics, but the former especially needs to be measured at document level. Appropriate terminology is vital for most domains. Outputs need to be vetted for unwanted bias. Literary and other creative texts require other criteria to be used (e.g. creativity, appropriateness of fictive dialogue, etc.).
- P22 - biases - user experience - no hallucinations - Skopos
- P23 - - The basic fluency and adequacy criteria - Is the information usable for specific contexts. It seems that most evaluation focuses solely on linguistic quality, but it would be important to also evaluate whether raw MT is usable in some situations. For example, is the information patent or law professionals get from raw MT sufficient for them making judgments about the importance and relevance of that information? This is a common and growing use case, but I haven't seen much research that tests its viability