Measuring Machine Translation User Experience (MTUX): A Comparison between AttrakDiff and User Experience Questionnaire

Vicent Briva-Iglesias
SFI CRT D-REAL, SALIS
Dublin City University
vicent.brivaiglesias2@mail.dcu.ie

Sharon O'Brien
SALIS
Dublin City University
sharon.obrien@dcu.ie

Abstract

Perceptions and experiences of machine translation (MT) users before, during, and after their interaction with MT systems, products or services has been overlooked both in academia and in industry. Traditionally, the focus has been on productivity and quality, often neglecting the human factor. We propose the concept of Machine Translation User Experience (MTUX) for assessing, evaluating, and getting further information about the user experiences of people interacting with MT. By conducting a human-computer interaction (HCI)-based study with 15 professional translators, we present a methodological paper in which we analyse which is the best method for measuring MTUX, and conclude by suggesting the use of the User Experience Questionnaire (UEQ). The measurement of MTUX will help every stakeholder in the MT industry - developers will be able to identify pain points for the users and solve them in the development process, resulting in better MTUX and higher adoption of MT systems or products by MT users.

1 Introduction

Recently, artificial intelligence has captured the attention of many stakeholders in our society, not only in specialised academic journals and conferences, but also among laypeople (Fast and Horvitz 2016).

Large language models have driven technological breakthroughs, and the state-of-the-art has evolved mainly through training bigger and bigger models, with more parameters, more training time, and ultimately more computational resources (Brown et al. 2020). Research in language technologies has become a race to see who owns and releases the biggest language model (Roose 2023). This has also provoked the reaction of academics who reflect on language technology research from a socio-technical perspective, promoting a move to a more human-centered development of such language technologies (Bender & Gebru et al. 2021), which goes beyond 'human in the loop' concepts.

In the language services or Translation Studies domains, MT is a technology that has had significant impact in the past few years, and its adoption and implementation in workflows has provoked some rejection from professional translators (Cadwell, O'Brien, and Teixeira 2018). Many professional users feel that their needs have not been considered in the development and deployment of these technologies, and have therefore felt dehumanised, commodified, with an accompanying loss of agency and status (Fırat 2021; Moorkens 2020). This results in a lack of acceptance and trust in these technologies, which is usually not a rejection of the technology, but a veto on the way in which MT is applied and used (Vieira 2020).

Human factors such as users' perceptions or experiences of MT as a tool that facilitates multilingual communication - regardless of whether we are talking about professional translators or other types of users - have often been overlooked. The focus of research has been on the quality and productivity benefits of using these technologies (Moorkens et al. 2018), neglecting human satisfaction and resulting experiences of such human-computer interaction. This paper aims to fill a gap in the literature by proposing the concept **Machine Translation User Experience (MTUX)** and recommending its application in language technology research and development processes to create better, user-centered language technology products, which would result in improved human-computer interactions. We first present the related work, followed by the definition of the term MTUX and the methodology used to discern the best method for

evaluating MTUX in multilingual communication processes.

2 Related Work

Since the emergence of MT, academia and industry have analysed its impact and implications for translation processes and multilingual communication (Briva-Iglesias 2023).

The focus of research has been on professional translators. Typically, attention has focused on the speed of production for translation (or productivity) through post-editing against the productivity without MT assistance (Jia, Carl, and Wang 2019). It has been shown that post-editing, in many situations, makes it possible to be more productive than translating without MT support (Sánchez-Gijón, Moorkens, and Way 2019). Hence, the introduction and adoption of MT in industry workflows to meet more agile, fast and urgent translation and/or localisation processes (ELIS 2022).

Some attention has also been paid to translation quality: Does the use of MT affect the final quality of a translation? Are translations done through post-editing worse than translations done directly by humans and without any MT intervention? Guerberof Arenas (2014), for example, reported in an experiment with 24 professional translators that there were no statistically significant differences in translation quality of texts produced with MT output against texts produced without MT assistance.

Nevertheless, the study of the perceptions and considerations that users have about their interaction with MT and new language technologies is scarce. Some experiments dealing with these topics have only been disseminated in a superficial, descriptive way. For instance, Etchegoyhen et al. (2018) analysed with a 4-point Likert scale what professional translators thought of post-editing in a subtitling workflow. More extensive consideration was undertaken by Pérez-Macías, Ramos, and Rico (2020), who studied the perceptions of professional translators towards MT in the migratory context. Rossi and Chevrot (2019) also looked at the perceptions of MT from translators from the European Commission. Other research has also focused on what lay users of MT think of such language technologies, like that of Nurminen and Papula (2018), where results suggested that lay users find MT useful and tend to use it for gisting and assimilation purposes.

Additional research has even reported that users' perceptions of language technologies, such as the perception that MT is a threat to their profession, or the level of trust they have in MT, have a strong correlation with the final translation quality in a professional setting (Briva-Iglesias, O'Brien, and Cowan, Forthcoming). This demonstrates the importance of considering users' perceptions when interacting with technologies, as perceptions can have direct correlation or association with final translation quality.

Besides, it is important to note that there has been no specific action to collect perceptions from previous research and introduce this human feedback into the process of developing, updating or improving new language technologies, since, as we have mentioned above, these new technological breakthroughs have been especially technical, but not sociotechnical, forgetting the human factor in multilingual communication (Olohan 2011). By presenting the concept of MTUX, we intend to suggest a solution to this problem.

3 Machine Translation User Experience (MTUX)

Nowadays, the close relationship between people and technologies allows us to say that multilingual communication can be seen as a form of human-computer interaction in many instances. We are not only talking about professional translators who use technologies in the performance of their daily tasks. We can also include a user who does not know a language and wants to understand a text by using an online MT system for assimilation purposes, or because they want to share this information in their own language with someone else. It is therefore key to understand and know how these different types of human-MT interactions work.

Human-Computer Interaction (HCI) focuses on analysing the interactions of people with different systems, products or technological tools (Dix 2010). Large technological companies typically have entire teams dedicated to usability or user experience (UX), with the aim of improving the experiences of users when interacting with tools and thus achieving an expected end result. This expected goal may be to achieve a higher customer conversion, for example.

However, in the field of multilingual communication, Translation Studies, and MT, the inclusion of HCI methods, among which we can find the study of human and subjective factors, has been largely neglected. The small number of studies are described below.

In a controlled evaluation, Läubli et al. (2020) examined whether the way source and target segments were presented had any effect on productivity and error detection. They concluded that a segment-by-segment (top-bottom) presentation gave better results than a side-by-side segment presentation. Paradoxically, most current CAT tools

still use side-by-side segment presentation. O'Brien et al. (2017) studied different functionalities of CAT tools from a HCI-perspective, and found some features that irritated professional translators and increased cognitive friction. Consequently, they made a series of recommendations suggesting that technology tool developers should work with users to implement improvements.

From another point of view, some first steps have tried to address this lack of HCI methods in Translation Studies and MT by introducing more transversal methodologies and methods. An example is the work conducted by Guerberof Arenas, Moorkens, and O'Brien (2021), who introduced a usability questionnaire to assess the impact of translation modality on what the final readers of translated text thought, as well as to devise whether they could perform different tasks with the different texts. Another interesting work was conducted by Koponen et al. (2020), who analysed the experiences of subtitlers when using MT and used the User Experience Questionnaire (UEQ) developed by Laugwitz et al. (2018) to measure UX. Karakanta et al. (2022) conducted a similar study, replicating the methodology of Koponen and colleagues, but with a bigger number of subtitlers and focusing on automatic subtitling. Both studies lead to the conclusion that subtitlers' experiences of using MT in subtitling or automated subtitling ranged from neutral to slightly positive. In a similar vein, Briva-Iglesias, O'Brien, and Cowan (2023) analysed whether traditional or interactive post-editing had any effect on the UX of professional translators or the resulting quality and productivity after such an interaction, concluding that the interactive post-editing modality caused a statistically significantly higher UX than traditional post-editing.

Going back to Koponen and colleagues' research, they made a modification of the validated UEQ to adapt it to the post-editing task, but no further analysis of consistency, validity or reliability was carried out. Moreover, only experiences during interaction with the tool were analysed, forgetting about pre- and post-task perceptions. This exclusion of elements may be problematic, as we may lose information from some crucial elements in the human-computer interaction.

By considering the above analysis of literature, it is clear that both academia and industry have focused on studying the *usability* of MT, which, if we follow the definition of this concept provided by the *ISO 9241-11:2018 on Ergonomics of human-system interaction*, is "[the] extent to which a system [...] can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO 2018). In the field of HCI, *usability* is a fragment of a much broader and currently relevant concept, *user experience* (or UX), which according to the same ISO standard above, is "[a] person's perceptions and responses resulting from the use and/or anticipated use of a system" (ISO 2018).

Therefore, we propose that pre-, during-, and post-task perceptions should be considered when assessing MTUX. We believe that further methodological consideration of MTUX is needed at present, as it would help all stakeholders involved in the fields of Translation Studies, the language services industry, the MT domain or the multilingual communication world.

In MT studies, there is little literature or research on the analysis of user experiences when interacting with MT. Why is this the case when MT is so relevant today? Why is the focus on training larger and larger language models and not on improving the user experiences of the systems? Or, alternatively, why are we not paying attention to what the needs of specific users are in order to adapt and personalise these technologies to users' needs? Our supposition is that developers of MT systems are concerned about a particular aspect of quality, normally calculated via BLEU scores or some variant, which is driven by MT system 'competitions', but that this has caused a rather narrow focus on system performance that assumes if the output is of good quality, all users of the system will be satisfied. However, this is a simplistic and untested hypothesis, especially seeing as MT systems have highly variable performance across different languages, text types, use cases and contexts.

Our aim in this paper is to discover the best methodology for analysing MTUX in a way that can be applied to the full spectrum of MT users, and that allows us to:

- Know what MT users experience when interacting with MT systems or, in other words, evaluating their MTUX.
- Discover the positive aspects that make the interaction with the system and the resulting MTUX satisfactory and positive (if applicable), with the aim of maintaining or enhancing them in the design or development stages.
- Discover the negative aspects that make the interaction with the system and the resulting MTUX unsatisfactory and negative (if applicable), with the aim of finding weaknesses in the system development and/or design step and thus taking into account the perceptions of real users in the development or updating of the systems.
- Adapt the tools for the different types of users who may use them: professional translators, people who do not know a language and use MT for

assimilation purposes, companies using MT for dissemination purposes or users of MT for foreign language learning, among many other scenarios.

Therefore, we propose the concept ***Machine Translation User Experience (MTUX)*** as "[a] person's perceptions and responses resulting from the use and/or anticipated use of MT". From this definition, we place a special emphasis in "resulting from", but also in "anticipated use". We consider that both pre-, during-, and post-task perceptions and experiences related to the interaction of a person with an MT system, product or tool should be equally considered.

Our suggestion is that MTUX should be used both in the Translation Studies sector to analyse what professional translators experience in their work according to their domain (translators specialised in legal texts will have different experiences and/or needs compared with subtitlers), as well as to discover what other MT users feel when interacting with MT (such as an academic with an L1 other than English who writes in their L1 and then translates the text with MT). We acknowledge these are not the only use-case scenarios where MTUX should be studied and analysed, but just some examples.

Moreover, MTUX is also crucial in technology development, as there should be a symbiosis and collaboration between the MT and the language technology sector to introduce feedback from actual users in order to carry out updates, modifications or changes in the tools that have an impact and a real repercussion on the final MTUX. This will become more and more important as we see MT becoming further embedded into other technologies like, for example, social media or educational technology tools and increased use of multimodal MT.

It would also allow for personalising technological tools to each use case according to the user, with their subsequent adoption and better reception among the community for which such personalisation is intended (O'Brien and Conlan 2018).

4 Methodology

In HCI, there has been substantial discussion about the methodology for measuring UX, and different methods have been proposed depending on the objective of each researcher or study (Obrist, Roto, and Väänänen-Vainio-Mattila 2009). Some examples put the attention on the Hedonic Quality (HQ) of a product, and pay closer attention to emotions, hedonic elements or sensations (Hassenzahl, Beu, and Burmester 2001), while others have focused on the Pragmatic Quality (PQ) of a product, paying closer attention to a mix of subjective and pragmatic elements (Vermeeren et al. 2010). However, the conclusion that has been reached is that questionnaires are the tool that best collects this type of data, and there are different questionnaires that are most commonly used in terms of UX in the HCI world, specifically AttrakDiff and UEQ (Law et al. 2009).

Therefore, when measuring MTUX, we need to have our goals and aims clear to be able to choose the most appropriate method, so that every stakeholder involved with MT can benefit from the results of MTUX evaluation, regardless of whether we are talking about professional translators, language service providers or lay users of MT. Thus, we consider that, when assessing MTUX, our objective must be twofold:

- On the one hand, that the MTUX results that we obtain are appropriate for analysing the interaction of people with MT, and that they reflect in a real way the needs, preferences and opinions that the user has of their interaction with the system or product being analysed.
- On the other hand, that these results in MTUX are not just theoretical and hedonic, but also pragmatic, since only obtaining subjective results that do not entail productivity or pragmatic effects would not be very viable nor feasible in today's industry, where economics and productivity are essential.

4.1 Questionnaires

AttrakDiff (Hassenzahl, Burmester, and Koller 2003) consists of 28 pairs of opposing adjectives (e.g. "confusing-clear", "bad-good") to be assessed using a 7-point Likert scale just after interacting with a tool, product or system. AttrakDiff focuses on three different factors: Pragmatic Quality (7 items that focus on the ease of use of the system or tool), Hedonic Quality (14 items that focus on the creation of pleasurable experiences) and Attractiveness (7 items focusing on the overall experience resulting from the interaction). AttrakDiff has been used for purposes including, but not limited to, measuring UX when interacting with Augmented Reality displays (Kim and Yoo 2021) or analysing factors influencing the purchase of kitchenware (Bevan et al. 2016). AttrakDiff can be used to measure the UX of a single product, to compare multiple products, or to measure the differences in UX of a product before and after applying design updates. An online platform allows questionnaires to be created and sent to participants semi-automatically[1].

[1] AttrakDiff platform: https://www.attrakdiff.de/index-en.html

For comparison, we have used the User Experience Questionnaire (UEQ) (Laugwitz, Held, and Schrepp 2008). This second questionnaire consists of 26 pairs of opposing adjectives (e.g. "unattractive-attractive"), which are also to be evaluated on a 7-point Likert scale after interaction with the system, product or tool. UEQ also focuses on Attractiveness (6 items assessing the overall experience of the interaction), Pragmatic Quality (12 items, but divided in three different subfactors), and Hedonic Quality (8 items that are also divided in two subfactors). In UEQ, Pragmatic Quality is divided into Perspicuity (4 items focusing on the ease of use and learning the tool/product), Efficiency (4 items focusing on the efficiency and practicality of the product under analysis), and Dependability (4 items that analyse whether the user feels in control of the interaction). Hedonic Quality is divided into Stimulation (4 items focusing on whether the product is interesting and motivating) and Novelty (4 items measuring the degree of innovation of the system or product). Like AttrakDiff, UEQ can be used to measure UX after an interaction with a product, but also to compare UX after using different products. The authors have also developed a tool to facilitate data analysis using Excel that performs automatic statistical analysis of validity and reliability (Schrepp, Thomaschewski, and Hinderks 2017). UEQ has been used in multiple scenarios, such as in the UX evaluation of different web page designs (Schrepp, Hinderks, and Thomaschewski 2014).

4.2 Participants

We recruited 15 professional translators in the English-Spanish combination and asked them to translate legal texts in Lilt, a CAT tool that offers the possibility of translating via traditional post-editing and interactive post-editing workflows. In order to obtain different measurements of MTUX, the translators interacted with the tool on two consecutive days (4 different interactions). Thus, on the first day, translators worked one hour with traditional post-editing and one hour with interactive post-editing, and on the second day they did the same but with different texts. After each hour of interaction with a post-editing modality, they completed both the AttrakDiff and UEQ questionnaires. The display of the questionnaire items at each point were randomised with positive and negative poles for each item alternated to avoid any confounding order effects or response acquiescence.

4.3 Analyses performed

To compare the two questionnaires and their reliability, i.e. the consistency of the analysed factors between participants, every perception (4 AttrakDiff questionnaires of 28 items by 15 translators: 1680 perceptions; 4 UEQ questionnaires of 26 items by 15 translators: 1560 perceptions; total of 3240 perceptions) was collected and analysed in different ways.

First, we made a comparison of the items. As some of the opposite adjective pairs measured in both questionnaires were similar (and in some cases even identical), we extracted the items that were similar in both questionnaires to be able to discern which questionnaire was more appropriate and adequate for measuring MTUX by considering both Hedonic and Pragmatic Quality elements, and created Tables 1, 2, and 3. In Section 6, we discuss the similarities and differences between questionnaires more in depth by considering the two-fold objective that MTUX evaluation should achieve, stated in Section 4.

One of the most commonly used methodologies to measure the internal consistency of a test or scale is to calculate Cronbach's alpha coefficient (Cronbach 1951). This is a statistical test that gives a score between 0 and 1, and indicates whether the items of a test or questionnaire measure the same concept and whether there is a connection between the different items of the test. Thus, the higher the number, the more consistent or reliable the method of assessment or measurement. Although there are different degrees of interpretation, a Cronbach alpha above 0.7 is usually considered to indicate the robustness of a measurement method (Tavakol and Dennick 2011). In clinical cases where a patient's life may be at risk, this threshold of robustness is usually set at 0.9 (Ibid.). For our use case, a Cronbach alpha score above 0.7 would be sufficient and would indicate a high robustness of the method used. Thus, for calculating the Cronbach alpha, we only used the similar items in both questionnaires, shown in Table 1, against the different factors of the questionnaires (i.e. Attractiveness, Perspecuity, Efficiency, Stimulation and Novelty). This allowed us to compare the internal consistency of both questionnaires.

Finally, in order to better choose which is the best method to evaluate MTUX, we also ran a Bland-Altman statistical analysis (Bland and Altman 1999). This statistical method compares the mean difference of two quantitative measurements and places them within limits of agreement. Thus, by comparing the results of the two measurements, we can see whether the two methods offer the same measurement for a specific item, or whether the difference in measurement deviate largely between methods (Giavarina 2015).

5 Results

5.1 Item Comparison

After comparing the different elements in each questionnaire, we could find 20 items that were very similar (or identical) both in AttrakDiff and in UEQ.

Table 1 shows these similar terms side-by-side, while also including the factor in which the questionnaires included each of the items. These factors are relevant for calculating the Cronbach alpha.

No.	AttrakDiff Item	UEQ Item	Factor
1	cumbersome-straightforward	not understandable-understandable	Persp. (PQ)
2	unimaginative-creative	dull-creative	Nov. (HQ)
3	unruly-manageable	difficult to learn-easy to learn	Persp. (PQ)
4	cheap-premium	inferior-valuable	Stimul. (HQ)
5	dull-captivating	boring-exciting	Stimul. (HQ)
6	unpredictable-predictable	unpredictable-predictable	Depend. (PQ)
7	conventional-inventive	conventional-inventive	Nov. (HQ)
8	bad-good	bad-good	Attrac.
9	complicated-simple	complicated-easy	Persp. (PQ)
10	unpleasant-pleasant	unpleasant-pleasant	Attrac.
11	ordinary-novel	usual-leading edge	Nov. (HQ)
12	bold-cautious	not secure-secure	Depend. (PQ)
13	discouraging-motivating	demotivating-motivating	Stimul. (HQ)
14	confusing-clearly structured	confusing-clear	Persp. (PQ)
15	impractical-practical	impractical-practical	Effic. (PQ)
16	tacky-stylish	cluttered-organized	Effic. (PQ)
17	ugly-attractive	unattractive-attractive	Attrac.
18	separates me from people-brings me closer to people	unfriendly-friendly	Attrac.
19	conservative-innovative	conservative-innovative	Nov. (HQ)
20	disagreeable-likeable	unlikable-pleasing	Attrac.

Table 1. Similar items from AttrakDiff and UEQ

It is worth stressing that AttrakDiff had 28 items and UEQ 26 items, resulting in 8 and 6 items without a similar opposite adjective pair. Table 2 contains these orphan items from the AttrakDiff questionnaire, as well as their relevant factors. Most of these orphan items (5 out of 8) focus on Hedonic Quality, so they put the attention on whether the human-computer interaction is pleasurable for the person, thus giving more importance to emotional elements. There is only one orphan item at AttrakDiff that focuses on Pragmatic Quality.

No.	AttrakDiff Item	Factor
1	technical-human	Depend. (PQ)
2	unprofessional-professional	Stimul. (HQ)
3	unpresentable-presentable	Novelt. (HQ)
4	rejecting-inviting	Attractiv.
5	challenging-undemanding	Persp. (HQ)
6	alienating-integrating	Persp. (HQ)
7	isolating-connective	Persp. (HQ)
8	repelling-appealing	Attractiv.

Table 2. AttrakDiff items without a similar comparison at UEQ

Table 3, on the other hand, shows the orphan terms from the UEQ questionnaire that had no similar item in AttrakDiff. We can clearly see a difference here, as the case is completely the opposite if compared with Table 2. Four out of six orphan items in UEQ are assigned to Pragmatic Quality (therefore focusing more on practical elements), while there is only one focusing on Hedonic Quality.

No.	UEQ Item	Factor
1	annoying-enjoyable	Attrac.
2	not interesting-interesting	Stimul. (HQ)
3	inefficient-efficient	Effic. (PQ)
4	does not meet expectations-meets expectations	Depend. (PQ)
5	slow-fast	Effic. (PQ)
6	obstructive-supportive	Depend. (PQ)

Table 3. UEQ items without a similar comparison at AttrakDiff

5.2 Questionnaire Reliability

The reliability of each questionnaire (i.e. whether every person who completed the questionnaire was consistent with their answers for the different scales) can be observed and analysed through the Cronbach alpha results in Table 4.

Factor	AttrakDiff	UEQ
Attractiveness	0.80	**0.93**
Perspicuity (PQ)	0.10	**0.85**
Dependability (PQ)	0.03	**0.70**
Efficiency (PQ)	0.84	**0.84**
Stimulation (HQ)	0.77	**0.78**
Novelty (HQ)	**0.85**	0.71

Table 4. Cronbach alpha results per Factor and Questionnaire

From Table 4 we can determine that the Cronbach alpha is higher for UEQ in 5 out of the 6 factors analysed. The only exception is the case of Novelty, where AttrakDiff attains a Cronbach alpha of 0.85, and UEQ only attains a Cronbach alpha of 0.71.

Nevertheless, the most important part is that AttrakDiff obtains very feeble and poor reliability scores in Pragmatic Quality factors, specifically in Dependability (0.03) and Perspicuity (0.10). UEQ obtains a Cronbach alpha score of 0.70 and 0.85 in these two factors, respectively. This indicates that AttrakDiff fails to measure in a reliable and consistent way the pragmatic elements of MTUX. It is also worth stressing that every Cronbach alpha result in UEQ is over the 0.7 threshold, therefore the reliability and consistency of this questionnaire should be considered acceptable and robust for every factor analysed.

5.3 Agreement between Questionnaires

Finally, for the Bland-Altman plot, we have analysed the different data points. We have only included the data points originating from the items that we consider as equal in Table 1. Should both questionnaires measure the same for these items, every data point (or at least most of them) should be within the confidence intervals.

Thus, we have 20 similar item ratings from 15 translators for 4 interactions (2 traditional post-editing and 2 interactive post-editing) = 1200 results for each of the questionnaires. We calculated the difference of the measurements and extracted the mean (0.5) and the standard deviation (1.42). From this result, we established the confidence intervals, and created a Brand-Altman plot to see if the values were within the limits of agreement. If so, it would mean that the questionnaires are consistent and measure the same construct for the categories we have matched and compared.

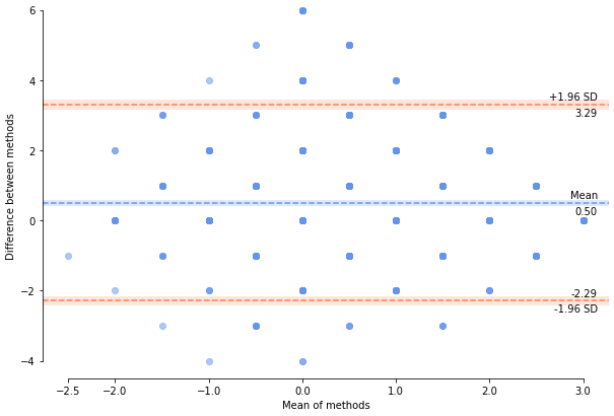

Figure 1: *Bland-Altman plot showing the differences in measurements between AttrakDiff and UEQ*

At a glance, we can see that although some of the data points lie between the 95% confidence intervals (red lines), there are still many data points beyond those interval lines. This means that the differences between the means of the items analysed were substantial.

If we analyse the data from Figure 1 more in depth and statistically, we can see that, from 1200 data points, 468 exceed the mean difference, thus being beyond the precision and confidence intervals. This means that 39% of the data points or perceptions were outside the expected confidence range, indicating that despite the constructs seemed to overlap for the two questionnaires they do not seem to measure the same thing consistently.

6 Discussion

By simply comparing items, we might conclude that 19 pairs of adjectives are very similar or identical between the two questionnaires. However, if we analyse the orphan items, we can see that AttrakDiff has a higher focus on the Hedonic Quality of the products evaluated, i.e. it is a questionnaire with a more emotional emphasis and in search of the user's pleasure. This questionnaire may be more appropriate for evaluating UX from a graphic design point of view, such as web page layout and functionality or applications whose objectives are creating hedonic pleasure for the user.

In contrast, UEQ puts more emphasis on the Pragmatic Quality of the product or system, and focuses more on efficiency, as we can see in the orphan pairs "inefficient-efficient" and "slow-fast", which are elements completely neglected in AttrakDiff. We suggest that this kind of adjective pair is very relevant for measuring MTUX, because whether or not MT users think the MT system, product or service helps them to be efficient or fast is

valuable information for analysing the interaction of users with MT. The relevance of this user perception is even more important if we are comparing two different ways of interacting with MT, such as traditional and interactive post-editing.

In the language services industry, a sector where productivity is vital (due to the fact that, if a translator works faster, this usually translates into higher profits for them personally or the company they work for), assessing Pragmatic Quality is a key element. Thus, we conclude that, in terms of items and adjective pairs, UEQ is more relevant for measuring MTUX because it combines both the hedonic and emotional views of users with more pragmatic and efficiency perceptions.

Item analysis is not the only fact that supports our preference for UEQ - the results of reliability and consistency between participants and factors through the Cronbach alpha coefficients also tip the balance towards the use of UEQ. In 5 out of 6 factors, the Cronbach alpha coefficient is higher in UEQ than in AttrakDiff. Novelty is the only factor where this is not the case, as AttrakDiff obtains a Cronbach alpha of 0.85 vs. 0.71 in UEQ. It is also worth stressing that AttrakDiff obtains very weak Cronbach alpha results in the factors related to the Pragmatic Quality of the product or system (0.10 in Perspicuity and 0.03 in Dependability; if compared with 0.85 and 0.70 in UEQ), which we already know to be of vital importance.

Finally, the Bland-Altman graph supports the results obtained by calculating the Cronbach alpha values and indicates that, although both questionnaires should report comparable measurements for similar or identical items, this is not the case. 39% of the perceptions and data points collected fall outside the 95% confidence intervals and, therefore, we can conclude that the two questionnaires do not measure these items in the same way.

Consequently, we believe that in a situation where both hedonic and pragmatic elements are of interest in the UX, as in the evaluation of MTUX, the appropriate method to use is the UEQ. In case we wanted to analyse any other type of UX more related with graphic design, for example, where the focus is more on the aesthetics of the tool or user pleasure, AttrakDiff may be a more appropriate choice.

7 Conclusions

In this article, we have identified a gap in the literature in the field of MT in multilingual communication processes: more attention needs to be paid to the users interacting with MT and not only to the productivity and quality of the tools. We believe that technical advances must go hand in hand with sociotechnical evaluation, which has been neglected to date (Olohan 2011).

We therefore present the concept of MTUX and explain the role that its adoption can play in the development of language technologies and especially MT, with the aim of creating sustainable and ethical language technologies.

For the first time, two of the most commonly used questionnaires in HCI for measuring UX have been applied to MT use in order to study MTUX. Data from 15 professional translators working in different iterations suggest that the best tool for measuring MTUX is the UEQ by considering both Pragmatic and Hedonic Quality criteria of the products or systems evaluated.

The adoption of MTUX analysis and study will help to create better experiences for any user of MT products or systems and will allow developers to include authentic human feedback in the design process in order to offer personalised tools according to the type of user. This will result in a wider adoption of language technologies or MT, and a better human-machine symbiosis that will bring us closer to Intelligence Augmentation (IA, as opposed to AI) (Sadiku and Musa 2021). By pursuing IA, we will be able to enhance and improve human skills and capabilities *thanks to* and *through* technology in a safe, secure, ethical, sustainable and human-centered way.

As for the limitations of the study, the evaluation of MTUX requires taking into account the pre-, during- and post-task perceptions of the users. In this paper we have addressed some methodological questions on how to measure MTUX by comparing two HCI-type questionnaires. We have not had the possibility of exploring how developers might apply results from the questionnaires or how those results could be triangulated with other measures, but this will be the focus of attention in the near future. In future work, we will introduce the Machine Translation User Experience Questionnaire (MTUXQ) to facilitate the analysis of all user perceptions related to MT interaction and semi-automate the statistical efforts that can be an initial barrier to the study of MTUX.

Acknowledgements

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224, and the 2021 competitive sponsorship of student's projects of the European Association for Machine Translation (EAMT).

References

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. ‘On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ’. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445922.

Bevan, Nigel, Zhengjie Liu, Cathy Barnes, Marc Hassenzahl, and Weijie Wei. 2016. ‘Comparison of Kansei Engineering and AttrakDiff to Evaluate Kitchen Products’. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2999–3005. CHI EA ’16. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2851581.2892407.

Bland, J Martin, and Douglas G Altman. 1999. ‘Measuring Agreement in Method Comparison Studies’. *Statistical Methods in Medical Research* 8 (2): 135–60. https://doi.org/10.1177/096228029900800204.

Briva-Iglesias, Vicent. 2023. ‘Translation Technologies Advancements: From Inception to the Automation Age’. In *La Família Humana: Perspectives Multidisciplinàries de La Investigació En Ciències Humanes i Socials*, Lucía Bellés-Calvera; María Pallarés-Renau, 137–52. Emergents 3. Publicacions de la Universitat Jaume I. Servei de Comunicació i Publicacions.

Briva-Iglesias, Vicent, Sharon O’Brien, and Benjamin R. Cowan. 2023. ‘The Impact of Traditional and Interactive Post-Editing on Machine Translation User Experience, Quality, and Productivity’. *Translation, Cognition & Behavior*.

———. Forthcoming. ‘Translators’ Pre-Task Perceptions of CAT Tools and MTPE, and Their Relationship with Translation Quality: Implications for Training’.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. ‘Language Models Are Few-Shot Learners’. arXiv. https://doi.org/10.48550/arXiv.2005.14165.

Cadwell, Patrick, Sharon O’Brien, and Carlos S. C. Teixeira. 2018. ‘Resistance and Accommodation: Factors for the (Non-) Adoption of Machine Translation among Professional Translators’. *Perspectives* 26 (3): 301–21. https://doi.org/10.1080/0907676X.2017.1337210.

Cronbach, Lee J. 1951. ‘Coefficient Alpha and the Internal Structure of Tests’. *Psychometrika* 16 (3): 297–334. https://doi.org/10.1007/BF02310555.

Dix, Alan. 2010. ‘Human–Computer Interaction: A Stable Discipline, a Nascent Science, and the Growth of the Long Tail’. *Interacting with Computers* 22 (1): 13–27. https://doi.org/10.1016/j.intcom.2009.11.007.

ELIS. 2022. ‘EUROPEAN LANGUAGE INDUSTRY SURVEY 2022’, 44.

Etchegoyhen, Thierry, Anna Fernández Torné, Andoni Azpeitia, Eva Martínez Garcia, and Anna Matamala. 2018. ‘Evaluating Domain Adaptation for Machine Translation Across Scenarios’. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L18-1002.

Fast, Ethan, and Eric Horvitz. 2016. ‘Long-Term Trends in the Public Perception of Artificial Intelligence’. arXiv. https://doi.org/10.48550/arXiv.1609.04904.

Fırat, Gökhan. 2021. ‘Uberization of Translation: Impacts on Working Conditions’. *The Journal of Internationalization and Localization* 8 (1): 48–75. https://doi.org/10.1075/jial.20006.fir.

Giavarina, Davide. 2015. ‘Understanding Bland Altman Analysis’. *Biochemia Medica* 25 (2): 141. https://doi.org/10.11613/BM.2015.015.

Guerberof Arenas, Ana. 2014. ‘Correlations between Productivity and Quality When Post-Editing in a Professional Context’. *Machine Translation* 28 (3): 165–86. https://doi.org/10.1007/s10590-014-9155-y.

Guerberof Arenas, Ana, Joss Moorkens, and Sharon O’Brien. 2021. ‘The Impact of Translation Modality on User Experience: An Eye-Tracking Study of the Microsoft Word User Interface’. *Machine Translation* 35 (2): 205–37. https://doi.org/10.1007/s10590-021-09267-z.

Hassenzahl, Marc, Andreas Beu, and Michael Burmester. 2001. ‘Engineering Joy’. *IEEE SOFTWARE*, 7.

Hassenzahl, Marc, Michael Burmester, and Franz Koller. 2003. ‘AttrakDiff: Ein Fragebogen Zur Messung Wahrgenommener Hedonischer Und Pragmatischer Qualität’. *Mensch & Computer 2003: Interaktion in Bewegung*, 187–96.

ISO. 2018. ‘ISO 9241-11:2018(En), Ergonomics of Human-System Interaction — Part 11: Usability: Definitions and Concepts’. 2018. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en.

Jia, Yanfang, Michael Carl, and Xiangling Wang. 2019. ‘How Does the Post-Editing of Neural Machine Translation Compare with from-Scratch Translation? A Product and Process Study’. *The Journal of Specialised Translation*, January, 60–86.

Karakanta, Alina, Luisa Bentivogli, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2022. ‘Post-Editing in Automatic Subtitling: A Subtitlers’ Perspective’. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 261–70. Ghent, Belgium: European Association for Machine Translation. https://aclanthology.org/2022.eamt-1.29.

Kim, Young Jin, and Hoon Sik Yoo. 2021. ‘Analysis of User Preference of AR Head-Up Display Using Attrakdiff’. In *Intelligent Human Computer Interaction*, edited by Madhusudan Singh, Dae-Ki Kang, Jong-Ha Lee, Uma Shanker Tiwary, Dhananjay Singh, and Wan-Young Chung, 335–45. Lecture Notes in Computer Science. Cham:

Springer International Publishing. https://doi.org/10.1007/978-3-030-68452-5_35.

Koponen, Maarit, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020. 'MT for Subtitling: Investigating Professional Translators' User Experience and Feedback'. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*, 79–92. Virtual: Association for Machine Translation in the Americas. https://aclanthology.org/2020.amta-pemdt.6.

Läubli, Samuel, Patrick Simianer, Joern Wuebker, Geza Kovacs, Rico Sennrich, and Spence Green. 2022. 'The Impact of Text Presentation on Translator Performance'. *Target. International Journal of Translation Studies* 34 (2). http://arxiv.org/abs/2011.05978.

Laugwitz, Bettina, Theo Held, and Martin Schrepp. 2008. 'Construction and Evaluation of a User Experience Questionnaire'. *International Journal of Interactive Multimedia and Artificial Intelligence* 4 (4): 76. https://doi.org/10.1007/978-3-540-89350-9_6.

Law, Effie Lai-Chong, Virpi Roto, Marc Hassenzahl, Arnold P.O.S. Vermeeren, and Joke Kort. 2009. 'Understanding, Scoping and Defining User Experience: A Survey Approach'. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 719–28. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1518701.1518813.

Moorkens, Joss. 2020. '"A Tiny Cog in a Large Machine"'. Ts.00019.Moo. John Benjamins Publishing Company. 2020. https://benjamins.com/catalog/ts.00019.moo.

Moorkens, Joss, Sheila Castilho, Federico Gaspari, and Stephen Doherty, eds. 2018. *Translation Quality Assessment: From Principles to Practice*. Vol. 1. Machine Translation: Technologies and Applications. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7.

Nurminen, Mary, and Niko Papula. 2018. 'Gist MT Users: A Snapshot of the Use and Users of One Online MT Tool'. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*. http://rua.ua.es/dspace/handle/10045/76049.

O'Brien, Sharon, and Owen Conlan. 2018. 'Moving towards Personalising Translation Technology'. In *Moving Boundaries in Translation Studies*, 81–97. Routledge.

O'Brien, Sharon, Maureen Ehrensberger-Dow, Marcel Hasler, and Megan Connolly. 2017. 'Irritating CAT Tool Features That Matter to Translators'. *Hermes: Journal of Language and Communication in Business* 56 (October): 145–62.

Obrist, Marianna, Virpi Roto, and Kaisa Väänänen-Vainio-Mattila. 2009. 'User Experience Evaluation: Do You Know Which Method to Use?' In *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, 2763–66. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1520340.1520401.

Olohan, Maeve. 2011. 'Translators and Translation Technology: The Dance of Agency'. *Translation Studies* 4 (3): 342–57. https://doi.org/10.1080/14781700.2011.589656.

Pérez-Macías, Lorena, María del Mar Sánchez Ramos, and Celia Rico. 2020. 'Study on the Usefulness of Machine Translation in the Migratory Context: Analysis of Translators' Perceptions'. *Open Linguistics* 6 (1): 68–76. https://doi.org/10.1515/opli-2020-0004.

Roose, Kevin. 2023. 'How ChatGPT Kicked Off an A.I. Arms Race'. *The New York Times*, 3 February 2023, sec. Technology. https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html.

Rossi, Caroline, and Jean-Pierre Chevrot. 2019. 'Uses and Perceptions of Machine Translation at the European Commission'. *The Journal of Specialised Translation (JoSTrans)*, January. https://halshs.archives-ouvertes.fr/halshs-01893120.

Sadiku, Matthew N. O., and Sarhan M. Musa. 2021. *A Primer on Multiple Intelligences*. Springer International Publishing.

Sánchez-Gijón, Pilar, Joss Moorkens, and Andy Way. 2019. 'Post-Editing Neural Machine Translation versus Translation Memory Segments'. *Machine Translation* 33 (1–2): 31–59. https://doi.org/10.1007/s10590-019-09232-x.

Schrepp, Martin, Andreas Hinderks, and Jörg Thomaschewski. 2014. *Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios*. https://doi.org/10.1007/978-3-319-07668-3_37.

Schrepp, Martin, Jörg Thomaschewski, and Andreas Hinderks. 2017. 'Construction of a Benchmark for the User Experience Questionnaire (UEQ)', June. https://doi.org/10.9781/ijimai.2017.445.

Tavakol, Mohsen, and Reg Dennick. 2011. 'Making Sense of Cronbach's Alpha'. *International Journal of Medical Education* 2 (June): 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd.

Vermeeren, Arnold P. O. S., Effie Lai-Chong Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. 2010. 'User Experience Evaluation Methods: Current State and Development Needs'. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, 521–30. NordiCHI '10. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1868914.1868973.

Vieira, Lucas Nunes. 2020. 'Automation Anxiety and Translators'. *Translation Studies* 13 (1): 1–21. https://doi.org/10.1080/14781700.2018.1543613.