Sheila Castilho & Sharon O'Brien (2017). Acceptability of machine-translated content: A multi-language evaluation by translators and end-users. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *16*, 120–136.

Acceptability of machine-translated content: A multi-language evaluation by translators and end-users

Sheila Castilho

ADAPT Centre, Dublin City University, Ireland
sheila.castilho@adaptcentre.ie

Sharon O'Brien

SALIS/ADAPT Centre, Dublin City University, Ireland
sharon.obrien@dcu.ie

As machine translation (MT) continues to be used increasingly in the translation industry, there is a corresponding increase in the need to understand MT quality and, in particular, its impact on end-users. To date, little work has been carried out to investigate the acceptability of MT output among end-users and, ultimately, how acceptable they find it. This article reports on research conducted to address that gap. End-users of instructional content machine-translated from English into German, Simplified Chinese and Japanese were engaged in a usability experiment. Part of this experiment involved giving feedback on the acceptability of raw machine-translated content and lightly post-edited (PE) versions of the same content. In addition, a quality review was carried out in collaboration with an industry partner and experienced translation quality reviewers. The translation quality-assessment (TQA) results from translators reflect the usability and satisfaction results by end-users insofar as the implementation of light PE both increased the usability and acceptability of the PE instructions and led to satisfaction being reported. Nonetheless, the raw MT content also received good scores, especially for terminology, country standards and spelling.

1. Introduction

Translation quality assessment (TQA) has long been an important topic in translation studies and, with the increased demand for translation on the industry side, the interest in TQA has also intensified (see Drugan, 2013, for a good overview). However, "theorists and professionals overwhelmingly agree there is no single objective way to measure quality" (Drugan, 2013, p. 35) and, therefore, the definition of translation quality and the various models that purport to measure it are still a source of intense disagreement between academia and industry – as well as within both sectors. This gap between industry and academia is even more problematic when machine translation (MT) is added to the equation. The lack of agreement on what is a "good" translation has also led to many approaches towards MT evaluation. Therefore, MT quality can be considered from a range of different perspectives and there is no single approach that suffices to address all evaluation purposes (Hovy, King, & Popescu-Belis, 2002).

In recent years, the increased demand for rapid translation has led to frequent use of MT in the translation industry. DePalma, Hegde, Pielmeier, Stewart and Hedge (2013) report results of a survey which found that more companies are adopting automatic translation systems in order to translate enterprise content. Castilho and O'Brien (2016) also identified an increase in the use of MT systems and even raw MT for technical documentation in the localization sector. According to Castilho and O'Brien's survey, the decision on whether to use MT appears to be guided by the following:

- when the user expectation of quality is not very high, for example, end-users of some technical documentation may have more tolerance for MT error, and
- a content type that was not translated before due to cost or effort involved may be a good candidate for MT only.

With this increase in MT usage, the need to assess the quality of such translations has also increased.

Even with recent advances in MT, it is still often assumed that raw MT output requires post-editing (PE) if it is to be used for more than gisting purposes; as a result, the practice of PE has received much attention. However, little is known about how *end-users* engage with raw MT text or PE text, or how usable this text is, in particular if users have to follow instructions and act on them. Few studies have attempted to identify the impact of different modes of translation (e.g. human translation (HT), raw MT output, light PE of MT, full PE of MT) on the end-user. Of the studies that do address this question (see section 1 for an overview), the main shortcomings are either that they tend not to deal with all aspects of usability or that they do not account for PE. Therefore, a more comprehensive study on usability and user satisfaction is called for in order to determine the levels of acceptability of MT and PE. The focus of this article is on comparing levels of acceptability among professional translators with those of end-users for both raw MT output and for post-edited MT output. For a discussion on what is understood by "acceptability", see section 2.2.

2. Related work

Even though TQA is a key topic in translation, academia and industry greatly differ on how to measure it. Whereas academia focuses on the theory and pedagogy of translation quality, TQA in the industry is limited mostly to somewhat arbitrary error typology models, where "one size fits all" (Lommel, Uszkoreit & Burchardt, 2014, p. 456). The introduction of MT systems has also contributed to the debate since the technology has introduced alternative ways of measuring quality, for example automatic metrics and PE effort. For Drugan (2013), when the issue of translation quality is considered, academia and industry are, essentially "pursuing different goals and asking different questions" (p. 37).

In one of two competing definitions of translation quality, Koby, Fields, Hague, Lommel and Melby (2014) acknowledge the gap between industry and academia and state the need for agreement on an objective way to measure translation quality:

> a quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account *end-user needs* [authors' emphasis] (p. 416).

One approach that considers the end-user consistently throughout the translation process by examining usability research from the perspective of translation is the user-centred translation approach (UCT) (Suojanen, Koskinen, & Tuominen, 2015). The UCT approach presents the perspective that clients, providers and translators should discuss the text specifications relative to the end-user and the purpose of the translation. The authors affirm that this process ultimately improves communication between clients, translators and end-users. Errors, especially translation mistakes in comparison to the source text (ST), are evaluated according to their relevance in terms of functionality and usability, and rather than searching for mistakes made by the translators, the usability team aims to eliminate problems that the end-users might encounter (Suojanen et al., 2015, p. 129). Inspired by the UCT approach, this research considers the end-user as the primary focus of the translation product. The research consists of a mixed-methods design that aims to analyse and evaluate different types of translation (namely MT translation, PE and a small amount of human translation) across three languages, with English as the ST, in a technical domain.

2.1 Post-editing

Although great advances in MT have been recorded in the past decade or so, PE is still the traditional means for achieving publication quality. Advances in MT have meant that PE has become a more common practice in the translation industry, which has led to much research on PE effort (e.g. Carl, Gutermuth, & Hansen-Schirra, 2015; De Almeida & O'Brien, 2010; Daems et al., 2015; Depraetere, 2010; Guerberof, 2014; Koponen, 2012; Lacruz & Shreve, 2014; Moorkens et al., 2015; O'Brien et al., 2012; O'Brien et al., 2013; O'Brien et al., 2014; Plitt & Masselot, 2010; Sousa, Aziz, & Specia, 2011; Specia, 2011). However, MT evaluation has not yet considered the acceptability of MT output with the end-user as the evaluator.

Accordingly, apart from the fact that very little research has been carried out on the impact of different modes of translation generally on the end-user, the impact of MT on end-users has been significantly under-researched. In fact,. Some examples include the works of Tomita, Shirai, Tsutsumi, Matsumura and Yoshikawa (1993), Fuji et al. (2001), Jones et al. (2005), Roturier (2006), Doherty and O'Brien (2012) and Stymne et al. (2012). The main shortcomings of these approaches to date are that they tend not to cover all the aspects of usability: whereas some of them (e.g., Tomita et al., 1993; Fuji et al., 2001) deal with the problem of comprehension by asking participants to answer comprehension questions after reading a task without considering task time, others present only a questionnaire without requiring any tasks to be performed (Roturier, 2006). The work of Doherty and O'Brien (2012) uses the ISO definition of usability in which usability is defined as effectiveness, efficiency and satisfaction in a specified context of use (ISO, 2002). However, this work does not consider post-edited text, focusing as it does only on the ST and raw MT output. Previous work carried out by the authors (Castilho, O'Brien, Alves, & O'Brien, 2014) showed that light PE improves the usability of texts translated from English into Brazilian Portuguese, therefore providing a natural hypothesis that light PE may also improve usability for languages that are more "challenging" for MT. In the case of this article, these more challenging languages are German, Japanese and Simplified Chinese. The industry partner who collaborated on this research also identified these languages as being of particular interest to them. There is, therefore, a need to determine the usability (and acceptability) of post-edited MT output among end-users.

2.2 Acceptability

The term "acceptability" has been used in various fields – linguistics, text linguistics, translation, and also in human–computer interaction (HCI) – to refer to the level of acceptance to the end-user (also called reader, user, receiver, etc.) of language, a text or a product.

For Chomsky (1969), acceptability is "a concept that belongs to the study of performance" – where performance relates to the "actual use of languages in concrete situations" (p. 4). In his view, acceptability is a matter of degree(s) and can be specified through various operational tests. For Beaugrande and Dressler (1981), acceptability is seen as "the text receiver's attitude that the set of occurrences should constitute a cohesive and coherent text having some use or relevance for the receiver, e.g. to acquire knowledge or provide co-operation in a plan" (p. 17); that is, the text should establish useful or relevant information that render it worth accepting. Moreover, the authors state that the attitudes of the text users "involve some tolerance toward disturbances of cohesion or coherence, as long as the purposeful nature of the communication is upheld" (Beaugrande & Dressler, 1981, p. 113). Puurtinen (1995) states that there are different types of acceptability and, therefore, "a more complex, flexible concept, which allows of [sic] such heterogeneity" is needed (p. 230). In this view, to be acceptable, a translation should conform to the target culture's norms and, therefore, to the reader's expectations.

Van Slype (1979) uses the concept of acceptability to assess the quality of MT and defines it as "a subjective assessment of the extent to which a translation is acceptable to its final user" (p. 92). Roturier (2006, p. 4) bases his definition of acceptability on Beaugrande and Dressler's fourth standard of textuality and outlines that

> acceptability does not only refer to the relevance a text has for its receiver, but also to the manner in which its textual characteristics are going to be accepted, tolerated, or rejected by its receivers", and, therefore, "users will find machine-translated documentation acceptable when they tolerate some of the textual disturbances caused by an MT process (Roturier, 2006, p. 157).

Acceptability is also associated with HCI. For Nielsen (1993), system acceptability "is the question of whether the system is good enough to satisfy all the needs and requirements of the users and other potential stakeholders, such as the users' clients and managers" (p. 24). In his model of system acceptability, Nielsen considers usability to be a narrow concern of the system acceptability model.

Acceptability in the research presented here is conceived, as per Puurtinen's and Chomsky's definitions, as a complex concept, consisting of various degrees that can be measured using a variety of methods. The notion of performance mentioned by Chomsky also complies with the view of this research, since the performance of participants when completing specific tasks is one of the methods used to operationalize acceptability. This research, therefore, focuses on acceptability as per Nielsen's acceptability model, where acceptability is composed of various categories. It complies with Beaugrande and Dressler's concept of acceptability, in which acceptability refers to the relevance of a text to its receiver, and agrees with Roturier's claim that acceptability also relates to the extent to which the characteristics of a text are "accepted, tolerated and rejected by its receiver" (Roturier, 2006, p. 4).

Finally, the study aims to measure the acceptability of MT instructional content via measures of usability, satisfaction and quality. Applying Nielsen's model to translation, a user will find a translation (raw MT or lightly post-edited MT in the case of this study) to be more acceptable if they are able to use the translation to perform tasks, regardless of any flaws it may contain. The user will be able to "tolerate some of the textual disturbances caused by an MT process" (Roturier, 2006, p. 157), or they will find the text less acceptable if the flaws in the translation affect their ability to use the text to some extent. Acceptability, then, is the over-arching concept being investigated in this study and it is influenced by factors such as usability, satisfaction and text quality.

2.2.1 Usability

The term "usability" is usually associated with HCI. For Suojanen, Koskinen and Tuominen (2015), usability is "ultimately about the *user's* relative experience of the success of use" [p. 14, emphasis in original]. Therefore, "almost any human activity can be studied from the point of view of usability" (Suojanen et al., 2015, p. 14) in which "we can all be perceived as users" (Suojanen et al., 2015, p. 33).

Usability was defined in an ISO standard, first in ISO 9241, and later in ISO/TR 16982:2002. In the latter standard, usability refers to "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO/TR 16982, 2002). Suojanen et al. (2015, p. 33), however, point out that the concept of usability is not broadly used in the field of translation and the few studies that research usability are predominantly limited to technical translation. Byrne (2014) defines usability as "the extent to which readers can read a text, understand its content and perform whatever task is required by the text quickly and accurately and the extent to which they find the experience difficult or easy" (p. 201).

This research focuses on usability as outlined by Suojanen et al. (2015), in which a product (a text, in the case of this research project) is considered usable "if users can typically use it in a satisfactory manner in the context for which it was intended" (p. 14). For carrying out the experiments, the ISO definition for usability was adopted and, therefore, the concepts of effectiveness, efficiency and satisfaction were applied.

2.2.2 Quality

As discussed in the introduction, translation quality is a slippery concept. The study presented here was conducted in collaboration with an industry partner, a multinational technology company that translates millions of words into many languages annually. As the main goal of this study was to measure the acceptability of translated enterprise content (MT vs PE), and the focus of the article is on both professional translators' assessments and those of end-users, it stands to reason that the translated texts need to be assessed by the regular method our industry partner applies to their content.

For this reason, quality is measured via a TQA questionnaire answered by professional translators. More details of the questionnaire are provided in section 2.4.

2.2.3. Satisfaction

As outlined previously, satisfaction has been identified as one of the elements of usability and is defined in the HCI field as how "pleasant it is to use the system" (Nielsen, 1993, p. 33). However, with the advance of usability research into other areas, the definition for satisfaction can also be seen to fit new needs, and, therefore, the term 'system' can be viewed as a synonym for 'product' which, in turn, could be seen as a synonym for translated text, in the case of this research: how pelasant it is to use translated text.

ISO has defined satisfaction as "freedom from discomfort, and positive attitudes towards the use of the product" (ISO 9241-11, 1998). In conformity with that, Rubin and Chisnell (2011, p. 4) have defined satisfaction as the "user's perceptions, feelings, and opinions of the product, usually captured through both written and oral questioning". Even though satisfaction may be seen as a subjective construct, its measurement allows us to establish a broad picture of the user's reaction to how well the product works (Byrne, 2006).

This research adopts the ISO definition as well as the Rubin and Chisnell definition of satisfaction presented above. Moreover, Byrne's view on satisfaction also informs the objectives of this research, the objective of which is to measure users' subjective reactions, opinions, perceptions and attitudes towards the translated texts. Satisfaction is measured by means of two approaches: (1) end-user opinions immediately after performing a task and (2) evaluations of translation quality by professional translators via a TQA questionnaire.

3. Method

3.1 Content

The corpus selected for the usability, quality and satisfaction experiments consists of Online Help articles from a software company for one specific piece of software, a spreadsheet application. However, what exactly Online Help means is open to interpretation. As Castilho and O'Brien (2016) show, labels for content types within the localization industry are fuzzy at best.

The Online Help articles both describe features of the spreadsheet application and give instructions on how to use such features; they are published on the company's website. The choice of content is motivated by several factors, including: (1) the easy access users have to this content online which would allow for a wide-scale survey on satisfaction; [1] (2) the willingness of the company to provide the content, and (3) the function of the content being instructional, which enables tasks to be created that users can perform during eye-tracking experiments.

Originally, the articles published online contained a few images of the software such as buttons, icons, etc.; however, as the objective of the experiment is to measure the usability of the text, some of the artwork was removed from the text. The authors made sure that any artwork essential to understanding the text was retained and that only complementary artwork was removed. Three English native speakers were asked to test the texts with the art removed. In total, only three images were left in the text, two in task 3 and one in task 6. Each task is listed below:

(1) Quickly change colours, fonts, and effects in your worksheet
(2) Change the font format for hyperlinks
(3) Format text in headers or footers
(4) Add a comment
(5) Apply conditional formatting with colour
(6) Insert an exploding pie chart0
(7) Insert a bar chart based on a pie chart

(8) Hide comments and their indicators.

Tasks 4 and 8 were created from the same article and were chosen because human-translated versions were available in the target languages (TLs) under study here (German (DE), Simplified Chinese (ZH) and Japanese (JP)). The human-translated content was factored in to enable additional comparisons to be made between human translation and PE and raw MT. However, this study reports only on the MT and PE content, and therefore, the results for tasks 4 and 8 will not be reported on here.

The MT system used to translate the instructions was Microsoft Translator, with a custom domain for end-user content, which was trained using the Microsoft Translator Hub. It is the production system used for the company's standard raw-MT publishing.

The PE of the MT output was performed by the company's language service-provider (LSP) using the guidelines developed by the company. Differently from full PE, for which the industry partner expects that the same level of quality is attained as for human translation, light PE was carried out if terminology did not conform to the client-specific glossary and if there were grammatical errors in the output. No edits were implemented for purely stylistic reasons.

3.2 Participants

3.2.1 Users

Fourteen native speakers of German volunteered for the study. The participants, six male and eight female, were aged between 21 and 51. Nine participants held a post-graduate degree and five had an undergraduate degree. Twenty-one native speakers of Simplified Chinese were recruited for the study. Their age range was from 20 to 39 years and nine were male and 12 female. Sixteen participants held a post-graduate degree and five had an undergraduate degree. Twenty-eight native speakers of Japanese volunteered for the study. The participants were between 18 and 56 years of age; 20 were female and eight male. Eight of these participants held a post-graduate degree; 11 held an undergraduate degree and nine participants were exchange students.

3.2.2 Translators

Eighteen moderators from the company's LSP (six for each language – DE, ZH and JP) participated. According to the LSP, a moderator is slightly different from a translator because moderators have "solid experience with reviewing and quality evaluation" (personal communication).

3.3 Tools

In collaboration with the industry partner, a spreadsheet application was selected from the office suite to be used as the software for the usability experiment. Because the office suite has more than 1.2 billion users, it was necessary to choose an application in which participants would be literate but not total experts so as to avoid previous experience as a confounding factor. The decision was also made to use the newest version of the software, 2013, as it was assumed that fewer people would have used that version. Previous experience with the tool was reported in a selection questionnaire, which showed that all the participants had already used previous versions of the software but very few had used the 2013 version.

The users' interaction with the spreadsheet, text and tasks was recorded with a Tobii T60XL wide-screen eye tracker – 24-inch monitor – with a 60 Hz sampling rate.[2] It has high screen resolution, allowing for studies of detailed stimuli, which is essential to this experiment since the participants need to have a clear view of all the spreadsheet features. The fixation filter used is the ClearView

Fixation Filter, set to 100 milliseconds for the fixation duration and 30 pixels/sample for the fixation radius. As the experiment contains text and pictures (the user interface – UI), the setup for mixed content stimuli was chosen.

3.4 Post-task questionnaire

A posts-task questionnaire in English was presented to the users after they had performed the usability experiments; it consisted of nine questions with a Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). For all the statements, except numbers 5 and 8, the higher score (5) indicates higher satisfaction (the opposite is true for statements 5 and 8):

1. The instructions were usable.
2. The instructions were comprehensible.
3. The instructions allowed me to complete all of the necessary tasks.
4. I was satisfied with the instructions provided.
5. The instructions could be improved upon.
6. I would consult these instructions again in the future.
7. I would be able to use the software again in the future without re-reading the instructions.
8. I would rather have seen the source (English) version of the instructions.
9. I would recommend the software to a friend or a colleague.

The post-task questionnaire is used to measure both the users' perception of quality and their satisfaction levels. The results for questions 7 and 9 are not presented in this article because both questions are more related to the software usability.

3.5 Translation quality assessment

The TQA questionnaire used in this research is a tailored version from the freely available KantanMT's framework.[3] The TQA instrument was designed in consultation with the industry partner. As a result, it consists of four error categories: adequacy, fluency, syntax and grammar (spelling and sentence structure), and style (terminology and country standards). The questionnaire also included one question addressing satisfaction. For adequacy and fluency, a 1–4 Likert scale was used, whereas for syntax and grammar, style, and satisfaction a 1–3 Likert scale was used.[4]

4. Results

This section reports results from the different experiments performed with translators (moderators) and end-users. As stated previously, the aim of this study is to understand the level of acceptability of raw MT and post-edited MT for professional translators and end-users. To this end, a two-way ANOVA and MANOVA are used. When calculating a two-way (M)ANOVA, it is possible to get pairwise comparisons that are based on the estimated marginal means, which are unweighted means that control for the effect of other variables. This is important when comparing the means of unequal sample sizes where each mean is taken into consideration in proportion to its sample size. The pairwise comparison tables display the factors and dependent variables combined in different ways, so that different interactions can be analysed. In this study, a two-way ANOVA is used to compare the effect of PE and Language on Satisfaction. A two-way MANOVA is used, for example, to compare the effect of PE and Language on Adequacy and Fluency. We also use repeated measures to assess both dependent variables as one; therefore, Adequacy and Fluency are considered to be the same variable. Repeated measures designs have the same participants measured in all conditions, that is, they compare the differences in mean scores under two or more different conditions (Howitt & Cramer, 2011, p. 230). It is important to highlight here that, owing to the exploratory nature of this research, assumption testing is not a concern since these analyses are exploratory (i.e. a new area of research). Moreover, the cut-off for significance used is 0.10% (Bernard, 2011, p. 485), which means that the confidence interval is 90%.

We first present results for quality as rated by the translators and users, followed by results for satisfaction as rated by both cohorts. Results for usability are reported along with quality as perceived by the end-users. Statistical significant results are shown with an asterisk (*).

4.1 Quality

4.1.1 Translators

Adequacy and fluency are presented together since they are both assessed via a 1–4 Likert scale, whereas syntax and grammar and style are presented together as they are assessed via a 1–3 Likert scale.

Adequacy and Fluency

A two-way MANOVA with repeated measures was conducted to compare whether TL (German – DE, Chinese – ZH, or Japanese – JP) and instruction type (i.e. raw MT or Post-Edited MT) have an effect on the translators' ratings of adequacy and fluency (see Table 1).

Table 1: Mean and SD for adequacy and fluency

Measure	Instructions type		Mean	SD
Adequacy	DE	MT	2.83	0.17
		PE	*3.67	0.17
	ZH	MT	2.39	0.10
		PE	*3.22	0.25
	JP	MT	2.78	0.09
		PE	*3.33	0.00
Fluency	DE	MT	1.72	0.39
		PE	*3.56	0.20
	ZH	MT	2.50	0.17
		PE	*3.05	0.48
	JP	MT	2.67	0.17
		PE	*3.61	0.35

The factor instruction type was found to have a statistically very significant effect on adequacy and fluency, where ($F(1, 12) = 100.86$, $p <.001$). This means that there is a statistically significant difference across the two PE levels, MT (M = 2.48, SE = 0.06), and PE (M = 3.40, SE = 0.06). Moreover, DE_PE (i.e., the post-edited German instructions), ZH_PE and JP_PE show very statistically significant higher ($p < 0.001$) ratings against their MT versions. This indicates that the lightly PE version of the instructions was a more adequate translation of the source text, according to the translator-moderators.

Regarding fluency, when comparing the PE instructions against their own MT version, a very statistically significant difference was also found for the DE_PE ($p < 0.001$), ZH_PE and JP_PE ($p < 0.005$) instructions. This indicates that the lightly post-edited version of the instructions was considered to be a more fluent translation of the source.

Syntax and grammar and style

A two-way MANOVA with repeated measures was conducted in order to compare whether TL and instruction type have an effect on syntax and grammar and style, which was measured on a 1–3 Likert scale (see Table 2).

Table 2: Mean and SD for syntax and grammar and style

Measure	Instructions type		Mean	SD
Spelling	DE	MT	2.44	0.54
		PE	2.72	0.25
	ZH	MT	2.67	0.44
		PE	2.67	0.44
	JP	MT	2.39	0.35
		PE	2.83	0.17
Sentence structure	DE	MT	1.11	0.10
		PE	*2.72	0.09
	ZH	MT	2.22	0.39
		PE	*2.61	0.54
	JP	MT	1.94	0.10
		PE	*2.50	0.29
Terminology	DE	MT	2.22	0.25
		PE	2.39	0.26
	ZH	MT	1.95	0.63
		PE	*2.67	0.58
	JP	MT	2.28	0.09
		PE	2.72	0.25
Country standards	DE	MT	2.55	0.25
		PE	*2.89	0.19
	ZH	MT	2.28	0.19
		PE	*2.72	0.35
	JP	MT	2.72	0.19
		PE	3.00	0.00

The factor instruction type was found to have a statistically significant effect on syntax and grammar and style, where ($F(1, 12) = 15.65$, $p <.005$). This means that when the factor instruction type is considered without distinctions between languages, there is a statistically significant differences across the two PE levels, MT (M = 2.23, SE = 0.08), and PE (M = 2.70, SE = 0.08).

Regarding spelling, the DE_PE and JP_PE show higher ratings for spelling. This indicates that the moderators found fewer spelling problems in the lightly post-edited version of the instructions. However, these results were not statistically significant. The Chinese language instructions scored the same for both PE and MT instructions, which indicates that, for spelling, both the raw MT and the light PE versions were equal.

For sentence structure, all the post-edited instructions show higher ratings when compared to their MT versions. This indicates that the lightly post-edited version of the instructions shows fewer sentence structure problems; however, these results were statistically significant only for the DE_PE ($p < 0.001$) and JP_PE instructions ($p < 0.05$).

Regarding terminology, when comparing the PE instructions against their own MT version, all the PE instructions show higher ratings. This indicates that the lightly PE versions of the instructions contain fewer terminology problems. These results were statistically significant only for the ZH_PE against the ZH_MT instructions at the $p < 0.05$ level. The German and Japanese language did not show statistically significant differences between the different instruction types.

For country standards, again, all the PE instructions show higher ratings. This indicates that the lightly post-edited versions of the instructions show fewer problems. These results were statistically significant for the DE_PE against the DE_MT instructions at the $p < 0.10$ level, and for the ZH_PE against the ZH_MT instructions at the $p < .05$ level. The Japanese language did not show statistically significant differences between the PE and MT instructions.

4.1.2 Users

To perform tasks using the different instructions, participants were divided into two groups; one group used the raw MT instructions and the other used the PE instructions. Neither group knew that the texts they were reading had been translated, as we did not want the participants to be biased. They were asked to perform six tasks (see section 2.1) and answer a post-task questionnaire regarding the use of those instructions.

Table 3: Mean and SD for goal completion

Measure	Instruction type		Mean	SD
Goal Completion	DE	MT	3.19	1.69
		PE	*5.08	1.28
	ZH	MT	2.86	0.95
		PE	*3.70	1.25
	JP	MT	3.12	1.14
		PE	*3.87	0.90

The factor instruction type was found to have a statistically very significant effect on goal completion, where $F(1, 57) = 14.13$, $p < 0.001$. This indicates that when the factor instruction type is considered without distinctions between languages, there is a statistical difference across the two PE levels, MT (M = 3.05, SE = 0.21) and PE (M = 4.21, SE = 0.22). The results for the usability experiment (Table 3) show that all the PE instructions show a statistically significant higher number of successfully completed tasks than their MT groups for all three of the TLs. The German participants who used the PE instructions were more successful than all the other groups.

Regarding efficiency, that is, the number of successful tasks completed (out of all possible tasks) when task time is taken into account, the factor instruction type was found to have a statistically very significant effect on efficiency, where $F(1, 57) = 17.79$, $p < 0.001$. This indicates that when the factor instruction type is considered without distinctions between languages, there is a significant difference across the two PE levels, MT (M = 4.40, SE = 0.45) and PE (M = 7.17, SE = 0.48) at the $p <.001$ level. Table 4 shows that those participants who used the PE instructions to perform the tasks were more efficient than those who used the MT instructions.

Table 4: Mean and SD for efficiency

Measure	Instruction type		Mean	SD
Efficiency	DE	MT	4.42	2.53
		PE	*8.00	2.74
	ZH	MT	3.65	1.62
		PE	*7.14	3.26
	JP	MT	5.13	2.26
		PE	6.39	2.53

However, this difference is not statistically significant for Japanese. Chinese seems to be most affected by the implementation of light PE, closely followed by German.

The first three questions of the post-task questionnaire (section 2.4) aimed at gathering users' perceptions of translation quality. Table 5 shows the results. The factor instruction type was found to have a statistically significant effect on PTQ, where $F(1, 56) = 4.59, p <.05$. This means that when the factor instruction type is considered without distinctions between languages, there is a statistically significant difference across the two PE levels, MT (M = 2.42, SE = 0.108) and PE (M = 2.75, SE = 0.11).

Table 5: Mean and SD for quality

Statements	Instruction type		Mean	SD
1 – Usable	DE	MT	2.63	1.19
		PE	3.17	0.75
	ZH	MT	2.4	1.07
		PE	*3.7	0.82
	JP	MT	2.54	1.05
		PE	3.07	1.22
2 - Comprehensible	DE	MT	2.13	0.99
		PE	3	1.41
	ZH	MT	2.6	0.84
		PE	*3.4	0.84
	JP	MT	2.69	1.11
		PE	2.8	1.08
3 - Effectiveness	DE	MT	1.88	0.99
		PE	*3	1.55
	ZH	MT	2.2	1.23
		PE	3	1.25
	JP	MT	2.08	1.04
		PE	2.4	0.83

For statement 1 (the instructions were usable), all those who read the PE instructions gave higher ratings when compared to those who read the raw MT. This indicates that those participants who used the lightly PE instructions believed the instructions were more usable than those who used the raw MT instructions. However, these results were only statistically significant for the ZH_PE when compared to ZH_MT ($p < 0.05$).

Similarly to statement 1, for statement 2 (the instructions were comprehensible), although all the PE groups show higher ratings when compared to their MT groups, these results were statistically significant only for the ZH_PE when compared to ZH_MT at the ($p < 0.10$).

Regarding statement 3 (the instructions allowed me to complete all of the necessary tasks), the PE group gives higher ratings when compared to the MT group. This indicates that those participants who used the lightly PE instructions believed they were more effective than those who used the raw MT instructions. Interestingly, these results were statistically significant only for DE_PE when compared to DE_MT at ($p < 0.05$). This is because the DE_PE group was the most successful group, that is, they had the most number of successfully completed tasks of all the PE groups. A moderate significant difference was found for the ZH_PE when compared to the ZH_MT ($p < 0.11$).

4.3 Satisfaction

4.3.1 Translators

As mentioned previously, the translator-moderators were asked to assess the translated instructions regarding fluency, adequacy, grammar, style and satisfaction. The statement for satisfaction assessment was "I would be satisfied sending this sentence to be published" and consisted of a 3-point Likert scale where 1 = No, 2 = Somewhat and 3 = Yes.

A two-way ANOVA with repeated measures was conducted in order to compare whether language and PE have an effect on the statements regarding the satisfaction of the moderators. Table 6 shows the mean for each language. The factor instruction type was found to have a statistically very significant difference for satisfaction, where $F(1, 12) = 40.90$, $p < 0.001$. This means that when the factor instruction type is considered without distinctions between languages, there is a statistically significant difference across the two PE levels, MT (M = 1.29, SE = 0.09), and PE (M = 2.14, SE = 0.09).

Table 6: Mean and SD for translators' satisfaction

Measure	Instruction type		Mean	SD
Satisfaction	DE	MT	1.11	0.10
		PE	*2.44	0.10
	ZH	MT	1.11	0.10
		PE	*1.83	0.50
	JP	MT	1.67	0.34
		PE	*2.16	0.29

A pairwise comparison found that the DE_PE (p < 0.001), ZH_PE and JP_PE ($p < 0.05$) show statistically very significant higher ratings for satisfaction. This indicates that the moderators were more satisfied with sending for publication the lightly PE version of the instructions.

4.3.2 Users

A two-way MANOVA with repeated measures was conducted in order to determine whether TL and PE level have an effect on the statements in the post-task questionnaire, which was measured on a 1–5 Likert scale. Table 7 shows the mean and SD for each language.

Table 7: Mean and SD for users' satisfaction

Statements	Instruction type		Mean	SD
4 - Satisfaction	DE	MT	1.63	0.52
		PE	2	1.1
	ZH	MT	2.2	1.03
		PE	2.5	0.97
	JP	MT	2.08	0.95
		PE	2.67	1.35
5 -Improvement	DE	MT	1.13	0.35
		PE	1.33	0.82
	ZH	MT	1.5	0.53
		PE	2	1.15
	JP	MT	2	1.08
		PE	1.93	0.96
6 - Re-use	DE	MT	3.38	1.3
		PE	3.17	0.98
	ZH	MT	3.2	1.4
		PE	3.7	0.67
	JP	MT	3	1.22
		PE	3.6	1.06
8 - English version	DE	MT	1.63	0.74
		PE	*2.67	1.63
	ZH	MT	2	0.67
		PE	*3.2	1.4
	JP	MT	3	1.08
		PE	3.13	0.83

Regarding statement 4 (I was satisfied with the instructions provided), all the PE groups show higher ratings. This indicates that those participants using the lightly PE instructions were more satisfied than those using the MT instructions. However, these results were not statistically significant for any of the groups.

For statement 5 (the instructions could be improved upon), although the DE_PE and ZH_PE groups show a higher rating[5] when compared to the DE_MT and ZH_MT groups, none of the results were statistically significant. Interestingly, the JP_MT group shows slightly higher ratings for statement 5 when compared to their PE group. However, no statistically significant difference was found for this result.

For statement 6 (I would consult these instructions again in the future), ZH_PE and JP_PE show higher ratings. This indicates that those groups that used the lightly PE translated version of the instructions were more inclined to consult the instructions again. DE_PE, however, shows lower ratings for statement 6 when compared to the DE_MT groups, which indicates that for the German language, the group who used the raw MT version was more inclined to consult the instructions again. However, none of these results were statistically significant at the p >.10 level.

Finally, regarding statement 8 (I would rather have seen the source (English) version of the instructions), all the three PE groups (DE_PE, ZH_PE and JP_PE) show higher ratings[6] when compared to the DE_MT, ZH_MT and JP_MT groups. This indicates that those participants who used the raw MT version were more inclined to use the English version than the PE participants. These results were statistically significant for the ZH_PE compared to the ZH_MT group (p <.05) and DE_PE group compared to the DE_MT group (p < 0.10). The JP_PE group shows slightly higher ratings for statement 8 when compared to their MT group; however, this result was not statistically significant.

5. Discussion

5.1. Quality

With regard to the quality of PE and MT instructions rated by the translators, the PE instructions received higher scores than the MT instructions for adequacy, fluency and sentence structure – all differences with statistical significance. The PE instructions received higher scores in all the measures when compared to the MT instructions, apart from those for the Simplified Chinese language for spelling, which shows the same mean for PE and MT instructions. Furthermore, the results of the TQA reflect the usability and satisfaction results insofar as the implementation of PE increased the usability of the PE instructions and the satisfaction as perceived by users and moderators. Nonetheless, the raw MT versions also showed good scores, especially for terminology, country standards and spelling. It is not surprising that the PE instructions were rated higher for these aspects, but the results demonstrate this explicitly through various means of acceptability assessment, by both translators and end-users, something that has been rare in similar studies to date.

We consider that these results are possible because the industry partner has their own MT systems, which are trained on their own specific content type and would therefore be expected to produce good output for terminology, country standards and spelling. These results also reflect the usability results where the participants who used the MT instructions were still able to complete a good number of tasks even though they required more time to do so (see Table 3).

The results for the post-task satisfaction questionnaire presented to the users after they had performed tasks show that for the majority of the statements regarding perceived quality (statements 1, 2 and 3) the PE instructions were scored higher than the MT instructions, with the statistically significant differences between PE and MT being mostly observed with the Simplified Chinese language. The result for German regarding statement 3 is interesting since the German MT group was able to complete more tasks when compared to the other MT groups, yet the DE_MT instructions were not considered very usable by those participants who used them. Japanese is the only language that did not present statistically significant differences between the PE and MT groups for any of these statements.

5.2. Satisfaction

The results for the question "I would be satisfied sending this sentence to be published" presented to the translators confirmed that PE instructions were statistically more satisfactory for the moderators than the MT instructions for all the translated languages. These results confirm that the implementation of light PE increased the quality of the instructions, especially for the Chinese and German languages. These results correlate with previous assessments of usability and the post-task satisfaction questionnaire where the German and Simplified Chinese seem to be more affected by the implementation of PE.

Regarding the level of satisfaction perceived by the end-users, the PE groups of all languages score higher for statements 4 and 8 when compared to their respective MT group. These results confirm that there was a difference in the perceived satisfaction between the PE and the MT groups, especially for the Simplified Chinese. This in turn, correlates well with the usability results where the Simplified Chinese always shows differences between the PE and the MT groups for the MT instructions and its MT group generally scored lower than the other MT groups.

6. Conclusion

The results for usability, quality and satisfaction have demonstrated that light PE had a significant effect on acceptability, as defined in section 1.2, where the PE instructions presented higher levels of acceptability in comparison to the raw MT versions. Nonetheless, the raw MT versions were also still usable and the participants who used those versions of the instructions were able to perform several of the tasks given (see Table 3). This result is comparable to that of Doherty and O'Brien (2014) in which the raw MT versions were also deemed usable in real-world scenarios. These results are confirmed across two types of user: translators acting as quality moderators and end-users of the instructions.

Another factor that was found to have influenced acceptability is language. It has been shown that German and Simplified Chinese had greater levels of acceptability regarding their lightly PE versions compared to the raw MT versions. The findings were less clear-cut for the Japanese language. In summary, this study has shown that the implementation of light PE directly and positively influenced acceptability for German and Simplified Chinese, more so than for Japanese and, moreover, the findings of this research show that different languages have different thresholds of translation quality.

As discussed previously, even though there has been increased interest in measuring translation quality, there is no agreement on an objective way to measure translation quality and, in addition, the needs of the end-users of those translations are generally disregarded. The attempts made to move towards a more dynamic quality model (e.g. TAUS DQF, QTLaunchpad, QT21) are now taking into consideration different views of translation quality (DQF and MQM models)m including the view of end-users as suggested by the broad definition of translation by Koby et al. (2014) and the UCT approach. This study therefore corroborates these attempts by demonstrating that translation quality should empirically factor in end-user perceptions and ability to use content.

7. Future work

A further step would be a qualitative study on MT and PE errors in order to ascertain whether specific error types lead to lower levels of acceptability. Another step would be to apply this study to a different set of languages or domains, ranging from more to less 'challenging' for MT. In addition, it is essential to measure both the acceptability of MT and PE in different content types and the impact that acceptability has on business factors such as willingness to buy or recommend a product or service, or even company reputation or client satisfaction. Further investigation into the impact of different translation modes on specific content types, and in different scenarios of use, is therefore warranted.

References

Beaugrande, R., & Dressler, W. (1981*). Introduction to text linguistics*. New York: Longman.

Bernard, H. R. (2011). Research methods in anthropology: Qualitative and quantitative approaches. Plymouth, UK: Atlamira Press.

Byrne, J. (2006). Technical translation: Usability strategies for translating technical documentation. Dordrecht: Springer.

Byrne, J. (2014). Scientific and technical translation explained: A nuts and bolts guide for beginners. Abingdon: Routledge.

Carl, M., Gutermuth, S., & Hansen-Schirra, S. (2015). Post-editing machine translation: A usability test for professional translation settings. In A. Ferreira & J. W. Schwieter (Eds.), *Psycholinguistic and cognitive inquiries into translation and interpreting* (pp. 145–174). Amsterdam: John Benjamins.

Castilho, S., O'Brien, S., Alves, F., & O'Brien, M. (2014). Does post-editing increase usability?: A study with Brazilian Portuguese as target language. *Proceedings of the Seventeenth Annual Conference of the European Association for Machine Translation, 16–18 June 2014, Dubrovnik, Croatia, 183–190.*

Castilho, S., & O'Brien, S. (2016). Content profiling and translation scenarios. *The Journal of Internationalization and Localization, 3*(1), 18–37.

Chomsky, N. (1969). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Daems, J., Vandepitte, S., Hartsuiker, R., & Macken, L. (2015). The impact of machine translation error types on post-editing effort indicators. *Proceedings of the 4th Workshop on Post-Editing Technology and Practice*, *November 3, 2015, Miami, USA, 31–45.*

De Almeida, G., & O'Brien, S. (2010). Analysing post-editing performance: Correlations with years of translation experience In V. Hansen & F. Yvon (Eds.), *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, 27–28 May 2010, St. Raphaël, France.

DePalma, D. A., Hegde, V., Pielmeier, H., Stewart, R. G., & Hedge, V. (2013). *The language services market: 2013.* Lowell, MA: Common Sense Advisory.

Depraetere, I. (2010). What counts as useful advice in a university post-editing training context?: Report on a case study. *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, 27-28 May 2010, St. Raphaël, France.

Doherty, S., & O'Brien, S. (2014). Assessing the usability of raw machine translated output: A user-centered study using eye tracking. *International Journal of Human-Computer Interaction, 30*(1), 40–51.

Doherty, S., O'Brien, S., & Carl, M. (2010). Eye tracking as an MT evaluation technique. *Machine Translation, 24*(1), 1–13.

Drugan, J. (2013). Quality in professional translation: Assessment and improvement. London: Bloomsbury Academic.

Fuji, M., Hatanaka, N., Ito, E., Kamei, S., Kumai, H., Sukehiro, T., Yoshimi, T., & Isahara, H. (2001). Evaluation method for determining groups of users who find MT "useful". *Proceedings of the Machine Translation Summit VIII "Machine Translation in the Information Age", 18–22 September 2001, Santiago de Compostela, Spain, 103–108.*

Guerberof, A. A. (2014). Correlations between productivity and quality when postediting in a professional context. *Machine Translation, 28*(3–4), 165–186.

Hovy, E., King, M., & Popescu-Belis, A. (2002). Principles of context-based machine translation evaluation. *Machine Translation, 17*(1), 43–75.

Howitt, D., & Cramer, D. (2011). *Introduction to statistics in psychology*. Harlow: Prentice Hall.

ISO (1998). ISO 9241-11:1998. Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability. Geneva: International Organization for Standardization.

ISO (2002). ISO/TR 16982:2002 Ergonomics of human-system interaction – Usability methods supporting human-centred design. Geneva: International Organization for Standardization.

Jones, D., Gibson, E., Shen, W., Granoien, N., Herzog, M., Reynolds, D., & Weinstein, C. (2005). Measuring human readability of machine generated text: Three case studies in speech recognition and machine translation. *Proceedings of ICASSP '05 IEEE International Conference on Acoustics, Speech, and Signal Processing 2005 – Volume 5, 18–23 March 2005, Philadelphia, USA, 1009–1012.*

Koby, G. S., Fields, P., Hague, D., Lommel, A., & Melby, A. (2014). Defining translation quality. *Revista Tradumàtica: tecnologies de la traducció* [Online], 12, 413–420. Available from: https://ddd.uab.cat/pub/tradumatica/tradumatica_a2014n12/tradumatica_a2014n12p413.pdf [Accessed 02 June 2016].

Koponen, M. (2012). Comparing human perceptions of post-editing effort with post-editing operations. *Proceedings of the Seventh Workshop on Statistical Machine Translation, June 7–8, 2012, Montréal, Canada, 181–190.*

Lacruz, I., & Shreve, G. M. (2014). Pauses and cognitive effort in post-editing. In S. O'Brien, L. W. Balling, M. Carl, M. Simard, & L Specia (Eds.), *Post-editing of machine translation: Processes and applications* (pp. 246–272). Newcastle upon Tyne: Cambridge Scholars.

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional qualitymMetrics (MQM): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció* [Online], 12, 455–463. Available from: https://ddd.uab.cat/pub/tradumatica/tradumatica_a2014n12/tradumatica_a2014n12p455.pdf [Accessed 24 May 2016].

Moorkens, J., O'Brien, S., da Silva, I. A. L., de Lima Fonseca, N. B., & Alves, F. (2015). Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation, 29*(3), 267–284.

Nielsen, J. (1993). *Usability engineering*. Amsterdam: Morgan Kaufmann.

O'Brien, S., Simard, M., & Specia, L. (Eds.). (2012). Workshop on post-editing technology and practice (WPTP 2012). Conference of the Association for Machine Translation in the Americas (AMTA 2012). San Diego, CA, 28 October.

O'Brien, S., Simard, M., & Specia, L. (Eds.). (2013). Workshop on post-editing technology and practice (WPTP 2013). Machine Translation Summit XIV. Nice, 2–6 September.

O'Brien, S., Balling, L. W., Carl, M., Simard, M., & Specia, L. (Eds.). (2014). *Post-editing of machine translation: Processes and application*s. Newcastle upon Tyne: Cambridge Scholars.

Plitt, M., & F. Masselot. (2010). A productivity test of statistical machine translation postediting in a typical localcontext. *The Prague Bulletin of Mathematical Linguistics*, 93, 7–16.

Puurtinen, T. (1995). Linguistic acceptability in translated children's literature (Unpublished doctoral dissertation). University of Joensuu, Joensuu.

Roturier, J. (2006). An investigation into the impact of controlled English rules on the comprehensibility, usefulness and acceptability of machine-translated technical documentation for French and German users (Unpublished doctoral dissertation). Dublin City University, Dublin.

Rubin, J., & Chisnell, D. (2011). Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. Indianapolis, IN: Wiley.

Sousa, S. C., Aziz, W., & Specia, L. (2011). Assessing the post-editing effort for automatic and semi-automatic translations of dvd subtitles. *Proceedings of the International Conference Recent Advances in Natural Language Processing, 12-14 September, Hissar, Bulgaria, 97–103* .

Specia, L. (2011). Exploiting objective annotations for measuring translation post-editing effort. *Proceedings of the Fifteenth Annual Conference of the European Association for Machine Translation, 30–31 May, Leuven, Belgium, 73–80.*

Stymne, S., Danielsson, H., Bremin, S., Hu, H., Karlsson, J., Lillkull, A. P., & Wester, M. (2012). Eye tracking as a tool for machine translation error analysis In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (pp. 1121–1126), 23–25 May 2012, Istanbul, Turkey.

Suojanen, T., Koskinen, K., & Tuominen, T. (2015). *User-centered translation*. Abingdon: Routledge.

Tomita, M., Shirai, M., Tsutsumi, J., Matsumura, M., & Yoshikawa, Y. (1993). Evaluation of MT systems by TOEFL. *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation, July 14–16, 1993, Kyoto, Japan, 252–265*.

Van Slype, G. (1979). Critical study of methods for evaluating the quality of machine translation. Brussels: Bureau Marcel van Dijk.

[1] The wide-scale survey was carried out by publishing both raw MT and post-edited versions of the same content on the industry partner's website over periods of months during this study and collecting responses from real online users to the question: Was this information helpful? Contrasts were made between responses to the raw MT and to the post-edited articles. For reasons of space, we do not discuss these results in detail here.

[2] We do not report here on the eye-tracking (fixation) data collected.

[3] https://www.kantanmt.com/ [Last accessed 18 February 2016].

[4] The difference in Likert scales is due to the fact that the adequacy and fluency rating was tailored using KantanMT's quality review instrument, whereas syntax and grammar were tailored using the industry partner's instrument.

[5] Note that for statement 5 a low rating indicates that the participants thought the instructions needed improvement.

[6] Note that for statement 8 a low rating indicates that the participants would have preferred to see the English version of the instructions instead of the translated version they used.