Measuring Acceptability of Machine Translated Enterprise Content

Sheila Castilho Monteiro. de Sousa

Lic., M.A.

Thesis submitted for the degree of

Doctor of Philosophy

School of Applied Language and Intercultural Studies

Dublin City University

June 2016

Supervisor:

Dr. Sharon O'Brien

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:_____

(Candidate)

ID No.: _____

Date:_____

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Dr. Sharon O'Brien for the continuous support and guidance throughout this PhD. Her expertise, understanding and patience were decisive for this research and for that I am deeply thankful. I could not imagine having a better advisor and mentor.

I am extremely indebted to Dr. Mike Dillinger for all his assistance and support. His advice and brainstorming helped me to implement new ideas and look at my research from a different angle – obrigada! I would also like to thank Dag Schmidtke for his invaluable assistance and for providing the resources used in this study.

My deep gratitude goes to all the participants in this research who kindly accepted to participate in this experiment.

I have been lucky to have people around that did not hesitate to help me in times of need: Aurelie Sicard, Federico Gaspari, Patrick Cadwell and Joss Morkeens – thank you. Also, thanks to all my friends in DCU and outside for making this journey less tough and less rainy. Thank you everyone for all the help, the pep talks and the enjoyable karaoke nights.

I also would like to thank my friend John Tinsley for being my loyal 'tea buddy' and often reminding me that this was 'just a PhD'. Thank you, Johnnie, for all the laughter.

Three other special people that could not be left out are my girls Amelia, Virginia and Flor. Thank you for all the support and for helping to give my mind a break with our girls' night.

Finally, to my friends and family in Brazil, this has to be in my mother language: obrigada à minha família – minhãs irmãs Sandra e Rosana pelo apoio incondicional e pela confiança de que eu sempre saberia o que fazer; minhas sobrinhas Sandriny e Hannah pelas risadas nas muitas mensagens e fotos; e ao meu pai que sempre quis me ver 'doutora' – embora não seja bem o tipo de doutora que ele queria, mas acho que conta também. E para as minhas amigas companheiras de sempre: Carla, Adriana, Gerusa e Karine – obrigada por me ourvir, por me incluir e

por não esquecer. E, por fim, para os meus eternos amores que não estão mais comigo mas para sempre no meu coração: Penny e Sheldon. Te amo.

This PhD journey has impacted my life on so many levels. I am grateful for everything it has brought me, from all the hard times to all the *craic.*

Publications and Presentations from this Research

Publications:

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

Castilho, S. and O'Brien, S. 2016. Content Profiling and Translation Scenarios. *The Journal of Internationalization and Localization,* 3(1).

Castilho, S. and O'Brien, S. 2016. Evaluating the Impact of Light Post-Editing on Usability. *IN: Proceedings of the Tenth International Conference on Language Resources and Evaluation, 23-28 May 2016, Portorož, Slovenia*, pp.310–316.

Presentations:

Castilho, S. 2016. Interdisciplinary Approaches to MT Quality Evaluation. Unpublished workshop paper at: *Interdisciplinarity for Impact Workshop, 01 June 2016*, *Trinity College Dublin.* Dublin, Ireland.

Castilho, S. and O'Brien, S. Forthcoming. Machine translation: Cognitive Load and Satisfaction among End Users. Unpublished workshop paper at: *The 5th International Workshop on Translation Process Research (TPRW5), 01-03 December 2016*, *University of Graz.* Graz, Austria.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AOI	Area of Interest
CAT	Computer-Aided Translation
DE	German
EN	English
FC	Fixation Count
FD	Fixation Duration
HT	Human Translation
INST	Instruction window (AOI)
JP	Japanese
MT	Machine Translation
MTE	Machine Translation Evaluation
PE	Post-Editing
PEp	Professional Light Post-Editing
PEz	Post-Editing Zero
RQ	Research Question
SCP	Source Content Profiler
ST	Source Text
TQA	Translation Quality Assessment
TT	Target Text
UCT	User-Centered Translation
UI	User Interface
VC	Visit Count
VD	Visit Duration
ZH	Simplified Chinese

“The truth is [still] out there”
The X-Files

Abstract

Sheila Castilho M. de Sousa

Measuring Acceptability of Machine Translated Enterprise Content

This research measures end-user acceptability of machine-translated enterprise content. In cooperation with industry partners, the acceptability of machine translated, post-edited and human translated texts, as well as source text were measured using a user-centred translation approach (Suojanen, Koskinen and Tuominen 2015). The source language was English and the target languages German, Japanese and Simplified Chinese.

Even though translation quality assessment (TQA) is a key topic in the translation field, academia and industry greatly differ on how to measure quality. While academia is mostly concerned with the theory of translation quality, TQA in the industry is mostly performed by making use of arbitrary error typology models where "one size fits all". Both academia and industry greatly disregard the end user of those translations when assessing the translation quality and so, the acceptability of translated and un-translated content goes largely unmeasured. Measuring acceptability of translated text is important because it allows one to identify what impact the translation might have on the end user – the final readers of the translation. Different stakeholders will have different acceptability thresholds for different languages and content types; some will want high quality translation, others may make do with faster turnaround, lower quality, or may even prefer non-translated content compared with raw MT.

Acceptability is defined as usability, quality and satisfaction. Usability, in turn, is defined as effectiveness, efficiency in a specified context of use (ISO 2002) and is measured via tasks recorded using an eye tracker. Quality is evaluated via a TQA questionnaire answered by professional translators, and the source content is also evaluated via metrics such as readability and syntactic complexity. Satisfaction is measured via three different approaches: web survey, post-task questionnaire, and translators' ranking.

By measuring the acceptability of different post-editing levels for three target languages as well as the source content, this study aims to understand the different thresholds users may have regarding their tolerance to translation quality, taking into consideration the content type and language. Results show that the implementation of light post-editing directly and positively influences acceptability for German and Simplified Chinese languages, more so than for the Japanese language and, moreover, the findings of this research show that different languages have different thresholds for translation quality.

Chapter 1 – Introduction

Translation quality assessment has long been an important topic in translation studies and, with the increased demand for translation on the industry side, the interest in translation quality assessment (TQA) has also intensified. 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, is still a source of intense disagreement between academia and industry – as well as within both areas. While academia focuses on the theory and pedagogy of translation quality, the industry is more concerned with real-world needs and requirements.

This gap between industry and academia becomes more problematic when machine translation (MT) is added to the scenario. The lack of agreement on what is a 'good' translation has also led to many approaches for MT evaluation and 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 and Popescu-Belis 2002).

The increased demand for fast translation has led to frequent use of machine translation in the translation industry. DePalma et al. (2013) report results of a survey which found that more companies are adopting automatic translation systems in order to translate enterprise content (see Section 1.1.1 for more details). Castilho and O'Brien (2016) also identified an increase in the use of MT systems and even raw machine translation for technical documentation in the localisation sector (see Section 1.1.1). The decision on whether to use MT relies, to some extent, on the users' expectations of quality (for example, according to the industry partners who participated in this research, end users of technical documentation are expected to have higher tolerance for MT errors than users of marketing content); and whether it is a content type that was not translated before due to cost or effort, which could be a good candidate for MT only.

Even with these recent advances in MT, it is still often assumed that raw MT output requires post-editing if it is to be used for more than gisting purposes, and therefore, the practice of PE has received much attention (e.g. De Almeida and O'Brien 2010; Plitt and Masselot 2010; Sousa, Aziz and Specia 2011; Specia 2011; O'Brien et al. 2013; Lacruz and Shreve, 2014; Guerberof 2014; Moorkens et al. 2015; Daems et al. 2015; Carl, Gutermuth and Hansen-Schirra 2015, Koponen 2016). However, little is known about how *end users* engage with raw machine translated text or post-edited text, or how usable this text is, in particular if users have to follow instructions and act on them. Very little research has been carried out on the impact of different modes of translation (e.g. HT, raw MT output, light post-editing of MT, full post-editing of MT) on the end user, for example, the works of Tomita et al. (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 address all the aspects of usability: while some of them (e.g. Tomita et al. 1993; Fuji et al. 2001) address the problem of comprehension by asking participants to answer comprehension question after reading a task without considering task time, others present only a questionnaire, without any tasks to be performed (Roturier 2006). The work of Doherty and O'Brien (2012) uses the ISO's 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 account for post-editing. Therefore, a more comprehensive study on usability and user satisfaction is necessary in order to determine end-users' levels of tolerance (that is, the tolerance of the real readers of those texts) with regard to machine translation and post-editing.

This research draws on the user-centered translation (UCT) approach (Suojanen, Koskinen and Tuominen 2015) as its aim is to investigate the acceptability for end users of raw and post-edited MT. The UCT approach is heavily based on the concept of user-centred design (UCD) from usability research, where information about the user is brought into the software development process. For the UCT approach, the users have a central role in the production of the translation and their preferences should be given priority over the client's if there is a clash between the two. The approach describes concrete tools and methods that the

translator can use in taking the end user into account, and, regarding evaluation, UCT concentrates on envisioning types of processes that will produce a variety of successful translations to serve the needs of different audiences (Suojanen, Koskinen and Tuominen 2015, p.128). For this approach, "errors, especially translation mistakes in comparison to the source text, are evaluated according to their relevance in terms of functionality and usability […]" (ibid., p.129).

In order to identify the tolerance of end users for machine translated and post-edited texts, regarding the final product of translation, this research uses the concept of 'acceptability' as defined by Chomsky (1969) as a "matter of degree" that can be specified through various operational tests. It also draws on the work of Puurtinen (1995) who describes acceptability as a "complex concept" and on Nielsen's (1993) acceptability model, where acceptability is composed of various categories. The view of acceptability also borrows from De Beaugrande and Dressler's (1981) concept in which acceptability refers to the relevance of a text for its receiver, and from Roturier's (2006) view in which acceptability also relates to the extent to which the characteristics of a text are accepted, tolerated and rejected by its receiver (see Chapter 3 for further discussion). Therefore, acceptability is operationalised through the concepts of usability, quality and satisfaction and is addressed through the main research question:

RQ: What factors influence acceptability levels of a machine translated text for the end user?

It is hypothesised that three main factors may influence the acceptability levels of translated texts from English into German, Simplified Chinese and Japanese: Post-editing Level, Language and Source Content. This research question (RQ) is further broken down into specific questions for each of the factors, which are fully described in Chapter 3, Section 3.3. In order to test this hypothesis, a variety of complementary experiments is carried out to test usability, quality, and satisfaction in collaboration with an industry partner. This industrial collaboration

allowed for a strong ecologically valid scenario since the company was able to provide:

- the data for the experiments, i.e. the Online Help articles and the use of their spreadsheet software.
- the machine translated versions of the articles, which was done via their machine translation systems trained on their own corpus.
- the light post-edited version of the articles, where the light post-editing was performed by the company's translators, using the company's own guidelines.
- the data from the web survey displayed on the company's website. One point to highlight about the web survey is that, normally the company has just a few Online Help articles online which were completely raw machine translated and none that was light post-edited and, therefore, they have made an exception for this experiment allowing both sets of articles (raw MT and post-edited versions) to be published online at different points in time.
- the moderators who assessed the quality of the translations and who are experienced in doing so.

Thus, the collaboration with the industry partner was invaluable in this research.

As in Doherty and O'Brien (2012), usability is defined as effectiveness, efficiency and satisfaction (ISO 2002). In the view of this research, the translation product is considered usable "if users can typically use it in a satisfactory manner in the context for which it was intended" (Suojanen, Koskinen and Tuominen 2015, p.14) and the extent to which users find this experience difficult or easy (Byrne 2014). The usability experiments consist of participants performing tasks with the machine translated post-edited instructions and source instruction (as well as two HT instructions), where goal completion (effectiveness) and efficiency (task time, number of successful tasks divided by task time) are computed. Cognitive data is also gathered via an eye-tracker. Eye-tracking measures have become well established as indicators of cognitive effort (Rayner 1998; Radach, Kennedy and

Rayner 2004) and have been adopted by translation research as a technique in recent years.

Quality is measured via a TQA questionnaire for the machine translated and post-edited content (we also include two articles translated fully by human translators (HT) – for reasons explained in Chapter 1, Section 1.1.1). This is done because the translated texts used in this research need to be translated and also assessed by the regular method our industry partner applies to their content, thus, ensuring ecological validity of the quality experiment. Another aim of using a TQA questionnaire was to verify how fluent and adequate the MT, PE and HT instructions were according to professional translators (moderators in this case – see Section 4.3.1.2) and how satisfied they were with those translations. The source content (English) is also assessed for comprehension, readability, and complexity with the help of two text analysis tools.

Satisfaction is defined as the "freedom from discomfort, and positive attitudes towards the use of the product" (ISO 9241-11, 1998), and even though it may be seen as a subjective measure, it may help to establish a broad picture of the user's reaction to how well the product works (Byrne 2006). Satisfaction is measured via three different approaches: i) end-users' ratings for satisfaction via a web survey, ii) a post-task satisfaction questionnaire (performed after the usability experiment) and iii) moderators' ratings (the latter is just applied for the machine translated and post-edited content).

The remainder of this thesis is divided as follows: Chapter 2 – Translation Quality Assessment, provides a detailed review of relevant literature carried out in several disciplines, such as translation studies, machine translation, post-editing, usability research, eye-tracking and cognitive research. Subsequently, Chapter 3 presents the definitions of acceptability, usability, quality and satisfaction as well as the motivation that guided this research. The research questions are then explained along with the hypotheses for each one. Chapter 4 describes the methodology applied in this research and is divided into Source Content and Translated Content. The chapter also discusses the measures and statistical tests used to analyse the data collected. Chapter 5 and Chapter 6 present the results of the experiments starting with the usability and cognitive data (Chapter 5), followed by quality and

satisfaction (Chapter 6). Chapter 7 discusses the results and how the research questions were answered. Finally, Chapter 8 provides the conclusion of the research, the limitations of the study, the contributions that this research has provided to the field, as well as potential future work.

1.1 Motivation

This section describes the motivation for the present research. Considering that this research aims at evaluating the acceptability in terms of usability, quality and satisfaction of both machine and human translated content in different post-editing levels (raw MT and light PE) and also the source content, it is necessary to address the motivation for evaluating i) MT and PE, ii) acceptability, and iii) why this methodology is extended to the source content.

1.1.1 Why MT and PE?

Today's organisations are overwhelmed with the need to create a huge amount of content, faster, customised, and for numerous media platforms, in order to support their products. This increased demand for fast translation has allowed for machine translation to become frequent in the translation process. A recent survey (DePalma et al. 2013) on the current state of the language outsourcing localisation market suggests that more companies are adopting automatic translation systems in order to translate enterprise content. Using responses from over 1,000 suppliers in the language outsourcing market, DePalma et al. report on the percentage of LSPs that offer a given service or technology such as Translation (Human), Machine Translation Post-editing (MTPE), Translation Technology (which includes CAT tools) and others. According to this report, since 2011 the number of LSPs who offer MTPE has grown from 37.75% to 44.09%. HT went from 94.33% to 96.80% - indicating that the demand for HT is still high, while Translation Technology went from 33.02% to 40.88%. In the same year, DePalma and Sargent (2013) presented a report based on buyers of language services and MT technology via 108 respondents who use MT in their companies. They found that 88% of those companies have used MT for 1-10

years and the most cited reasons for using MT are: reducing cost; the need for speed; the desire to enter more markets; and the desire to provide better support to international customers. Reasons for not using MT include: linguistic quality, technical complexity, pricing models, lack of language support, etc. The authors also asked the participants how they see the quality level of MT systems. One per cent (1%) said that the quality is 'excellent'; 10% said it is 'good'; 66% 'fair'; 14% 'poor'; 3% 'horrible' and 6% say 'it depends'. Sixty per cent (60%) of the companies publish their MT output after some external or internal post-editing. Only 8% of the companies publish their MT output immediately. In general, MT output is rarely published without some kind of PE. When asked who they target with the MT output content, the participants mentioned the following external audiences - customers (62%), website visitors (40%), and prospects – potential clients (11%), as well as internal employees.

In order to identify how multinational companies with localisation needs are currently profiling content and how translation decisions are made based on this profiling, a survey was conducted by the author of this research (Castilho and O'Brien 2016), with professionals who participate in the decision making about content translation and localisation from six multinational companies. The results of this survey also identified an increase in the use of MT systems (or CAT+MT+PE), and even raw machine translation for technical documentation in the localisation sector. Figure 1:1 shows the findings of that survey regarding translation strategies according to the content types presented in the questionnaire and identified among those companies, and Table 1:1 shows translation strategies for content types that did not fall into the categories presented in the questionnaire. The decision on whether to use MT appears to be guided by the following: i) when the user expectation of quality is not very high, e.g., technical documentation is expected to have end users with more tolerance for MT errors; and ii) a content type that was not translated before due to cost or effort may be a good candidate for MT only.

With the increase in MT usage, the need to assess the quality of those translations has also increased. Recent efforts by the Translation Automation User Society (TAUS) showed that the translation industry is eager for a change in the error-based approaches to measure translation quality (as discussed in Chapter 2), for a framework that would account for variables such as content type, *end user requirements*, perishability, or mode of translation creation, that is, whether the translation is performed by a human translator with no help of CAT tools, MT system (raw or post-edited) or TM systems or even a combination of these (O'Brien 2012, p.55).

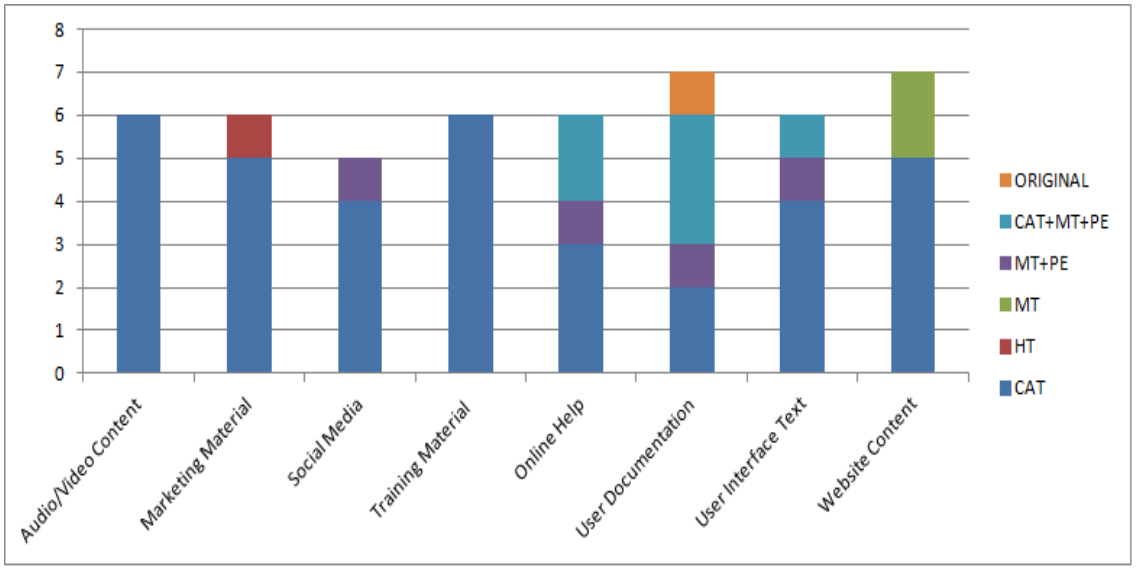


Figure 1:1 - Content Types vs. Translation Mode (Castilho and O'Brien 2016)

Content Type	Translation Technology	Strategy
Employee engagement survey	CAT	*moving to MT+PE
Internal announcements	CAT	
Support documentation	MT	*not translated before
Online knowledge base	MT	*not translated before
Legal texts	CAT	*sometimes translated only for specific countries
Surveys	CAT	
User generated and Industry generated	CAT+MT+PE	*experimenting
Sales training	CAT	*moving to MT+PE
Internal Sales tools	CAT	
Internal Training Material	CAT	
Metadata	CAT	
Templates	HT	*only because it is not TM/MT readable
Technical developer documentation	MT	*sometimes not translated

Table 1:1 - How content types not listed in questionnaire are translated (Castilho and O'Brien 2016)

Finally, the motivation for implementing light post-editing in this study (as opposed to full post-editing) comes from previous work by Doherty and O'Brien (2013), as discussed in Chapter 2, Section 2.7.1, that demonstrated that raw machine translation output can have a reasonable level of acceptability but that the levels of acceptability are higher for languages that are considered to be "easy" for MT (Spanish, in their case) in comparison to 'difficult' languages for MT (Japanese and German). Building on this, we hypothesise that by implementing light post-editing the levels of acceptability would be increased when compared to their MT versions. While these assumptions may seem obvious, it is important to reemphasise that very little empirical research has been carried out to test the acceptability of raw MT output, that the usefulness of so-called 'light' post-editing is also under-researched, and that there is a need for further testing of results such as those presented by Doherty and O'Brien, which was a small-scale study with different target languages and content. Moreover, a pilot study undertaken in November 2013 (see Section 4.1) found that light post-editing improves the usability of the texts translated from English into Brazilian Portuguese, thus providing a natural hypothesis that light post-editing may also improve the usability for languages that are more 'challenging' for MT. Another motivation for the implementation of light post-editing was that consultation with the industry partner

in this study resulted in a conclusion that light post-editing was of more interest to them since they expect full post-editing to produce quality that is indistinguishable from human translation.

The research also includes a small amount of human translation. The idea is to add HT as a control task, in order to verify whether there are differences in usability, quality and satisfaction when using HT when compared to the MT and PE texts across languages.

In light of what was presented above, the present research has identified a need for assessing MT and light PE (and HT) quality in a manner that is appropriate for the end user of those translations. The following Section presents the motivation for using acceptability as a measure for MT and PE quality.

1.1.2 Why Measure Acceptability?

This research considers acceptability to be composed of different constructs and is measured via usability, quality and satisfaction. The motivation for measuring acceptability of different post-editing levels (no post-editing, i.e. raw MT, and light post-editing) comes from the lack of studies on usability for MT (see details in Chapter 2). Even though one of the factors, quality, has been extensively studied – the usability of those translations, together with the satisfaction of end users when using the translation, go largely unmeasured.

1.1.3 Why Also Measure Acceptability of the Source Text?

The previously mentioned survey by the author (Castilho and O'Brien 2016) also found that source content evaluation does not follow a standard practice among companies and it is often ignored in the industry. That conclusion is also shared by Molnár (2012) who states that "TQA tends to be restricted to the target text […] as the final product of the translation process, with the ST [source text] being somehow neglected or even omitted from consideration" (ibid., p.61). Molnár affirms that a large portion of the source texts produced is written by non-professional authors who are often non-native speakers of English and that the

translators are the ones who have to handle them. The author conducts a survey in order to shed light on questions about how translators deal with those defective texts, i.e. to what extent translators can intervene; whether there are guidelines or any objective criteria to follow; what strategies translators adopt when facing faulty source content. The results of Molnár's survey show that the types of errors translators most frequently find in the source texts are spelling and punctuation (62%), stylistics (60%) and incomprehensibility (58%) among others. Interestingly, 72% of the translators consult the translation initiator before correcting a source text defect, and 15% of the translators always correct defects in the source text regardless of the time it requires, and only 3% trust the source text and do not expect any defects in it.

The issue of inconsistent or poor source language content is mentioned frequently by translators who have to make sense of ambiguous source language content and terminological or stylistic inconsistencies. Of course, the translation of repeated source language content has been catered for by the introduction of translation memory (TM) tools. Yet, TMs do not eradicate source language content issues, and can even store them for replication over many translation iterations (see discussion in Moorkens 2012). Poor and inconsistent source language content also contributes to poor quality machine translation (MT) output, which increases in turn the post-editing effort.

These results of Molnár's survey are interesting when contrasted with the previously mentioned survey conducted by the author of this research - which aimed at gathering information on how multinational companies with localisation needs are currently profiling content and how translation decisions are made, based on this profiling, information on source content strategies was also gathered (Castilho and O'Brien 2016). A summary of the findings will be presented here in order to give an overview of the industry practices that lead to the motivation for testing the source content.

Questions about authoring and source content were presented in the survey regarding i) guidelines; ii) cooperation with translation teams; iii) evaluation; and iv) end-user evaluation. The majority of the participants responded that they had in-house authors but a few of them have a small percentage of outsourced authors.

When asked if the company had any guidelines for authoring, all participants confirmed that they had some guidelines for source content, however, the participants claim that writing guidelines for authors is a very hard task because of divisions between groups inside the same company making it hard to set cross-divisional rules. Even though all the respondents confirmed that there is some kind of cooperation between the translation and authoring teams, they frequently report that the cooperation is between a small number of authoring teams only and that they are actively 'trying to bridge the gap' between both worlds.

The survey also tried to identify how source content evaluation takes place by investigating:

a) how the companies identify bad quality source text;

b) whether the source content is published before translation;

c) whether the feedback from translators is the factor that decides if the content is bad;

d) what happens to bad quality source; whether faulty source is sent back to the authoring team;

e) whether translators are expected to correct the source while translating

	a) How do you identify the bad quality of the source?	b) Is the source content published before translation?	c) Is the feedback from translators the factor that decided if the content is bad?	d) What happens to bad quality sources? Do you send it back to the authoring team?	e) Are translators expected to correct the source while translating?
Company B	Acrolinx	Published Simultaneously	One of the factors	Source is sometimes sent back Feedback is sent while translating	Translators should try to address the issue without changing the source This misaligns the TM matches in the future
Company C	Not done Translators point out issues (queries) and the queries are tracked for later analysis	No	Yes	Source is sometimes sent back (query system) If there is time or if the error is misleading to the user	Translators correct the translation but do not change the source This misaligns the TM matches in the future
Company D	Copy-editing	Big launches - simultaneously Lower priority - sometimes English may be first	One of the factors, but minor event	Source is sometimes sent back but just in case of severe problems (very rare) Copy-editors are supposed to correct.	Translators do not correct bad source
Company E	Automated validation checks	Published Simultaneously	One of the factors	A file cannot enter the translation process unless it is passed as valid by these tools	Translators do not correct bad source This would break one of the fundamentals of source control - translation management
Company F	Automated validation checks	Big launches - simultaneously	One of the factors	Source is sent back if it does not pass the automated validation checks	Translators do not correct bad source Translator may handle errors in the source with the translation and feedback is sent

Table 1:2 - Source Content Quality Assessment (Castilho and O'Brien 2016)

Table 1:2 presents a summary of the answers for source content evaluation. Regarding item a), three companies stated they use automated validation checks; copy-editing is used by one company and, one company does not do source evaluation before sending to translation. For item b) two companies said they

publish the content simultaneously while two others said they publish simultaneously if it is a big product launch. Otherwise, the English is published first, and one company said source content is always published first. When asked about item c), only one company answered that the feedback from translators is the factor that decides if the content is faulty. The other companies said translators' feedback is only one of the factors, as the preparation phase (automated or copy-editing) should identify most of the issues. Regarding item d), most of the companies said they 'sometimes send source back' and the reasons for that vary greatly. One company stated that the translation of the source starts a little after the source creation starts; therefore, creation and translation happen almost simultaneously. In this case, feedback from the translators is sent while translating. Others stated that bad quality source is sent back only when there is enough time or, even when time is an issue, if the source is misleading the user; it has to be sent back. Another company stated that source is sent back only in case of severe problems. However, they claim it is a very rare event as the copy-editors should correct those issues. The other two remaining companies said the source does not enter the translation process unless it is validated by the automated checks. Finally, when asked about item e), all companies said the translators should not correct the source, but they should handle any issues that may make it to the translation cycle. Again, the reasons for that vary greatly. While some companies declared that translators cannot change the source as "this misaligns the TM matches in the future" others stated that translators cannot correct the source because it is the copy-editors' job to do so. One company said translators are not supposed to correct source because this "would break one of the fundamentals of source control – translation management". And another stated that translators do not correct source and if any errors make it into the translation process, feedback is sent to the authoring team.

Regarding end-user evaluation, only one company said they do end-user evaluation for both product and content. It is clear that end-user evaluation is another point that seems to be underdeployed.

Source content can be problematic not just in commercial settings, but also in governmental entities, as is the case of the *European Commission's Directorate-*

General for Translation (DGT). In a project in the DGT in which the author of this research project was involved (Cadwell et al. 2016 [submitted]), it was identified that source content problems is a current issue in the translation process. The criteria for dealing with faulty source content in the DGT context, however, is more strict regarding translators correcting the source since it may imply changes to the legislation described in the texts they process. In the DGT context, translators contact the requester to try to understand the meaning intended in the source and maybe even help the requesters to correct the source text. Another interesting fact that arose in this study was that some of the writers/requesters are generally non-native speakers of English, which could contribute to problems in the source content, and, therefore, translators in the DGT feel they add unique value to the work carried out in the DGT by collaborating with requesters, national authorities, and each other to improve the quality of European Commission legislation. This is an interesting point that has also been researched by Vandepitte et al. (2010) who proposes a tutorial for both writers and translators with the goal of familiarising "professional communicators with the challenges that professional translators face when localizing the texts that communicators send them for translation" (p.58). The tutorial aimed at providing the audience with a perspective of translation projects from the translators' point of view.

From the evidence above, it is clear that source content can be problematic and still, source content is under-evaluated, which leads to translators having to deal with source problems and having to correct them. But how can machine translation interact with the requester/writer? Once faulty content is submitted to an MT system to be translated, the error will most likely to be propagated into the MT output. If post-editing is to be applied to that output, will the errors be dealt with? Therefore, this research assumes that measuring the acceptability of the source content is essential since it can shed light on whether the source contains any of those characteristics which, when translated by a machine translation system, could affect usability levels.

Chapter 2 – Literature Review

Translation Quality Assessment

In this chapter, an overview of translation quality assessment in the field of translation studies, the translation industry and machine translation is presented. Section 2.1 presents a brief introduction to the TQA approaches and introduces the main objectives of the chapter. Translation quality assessment practices both from a translation industry perspective (Section 2.3) and that of the translation studies field (Section 2.2) are analysed. A discussion about the gap between academia and industry is presented in Section 2.4, followed by a description of the user-centered translation approach. Section 2.6 introduces an overview of machine translation evaluation practices, including automatic metrics, human judgements and usability methodologies. Finally, Section 2.7 describes how eye-tracking methodologies have been used in cognitive research and how they have been adopted by the translation field to evaluate translation quality.

2.1 Introduction

The task of evaluating the quality of translation has raised debate and led to research among those interested in translation (namely, in translation studies, machine translation and in industry), and such interest has created a field known as translation quality assessment (TQA) (Secară 2005).

Even though TQA is a key topic in translation, academia and industry greatly differ on how to measure it. As Drugan outlines, “theorists and professionals overwhelmingly agree there is no single objective way to measure quality” (2013, p.35). While academia focuses on the theory and pedagogy of translation quality, TQA in the industry is mostly limited to somewhat arbitrary error typology models where “one size fits all” (Lommel, Uszkoreit and Burchardt 2014, p.456). The

introduction of machine translation systems has also contributed to the debate since the area has brought alternative ways of measuring quality, e.g. automatic metrics and post-editing effort. For Drugan, when the issue of translation quality is considered, academia and industry are, essentially, "pursuing different goals and asking different questions" (Drugan 2013, p.37). One thing that is common across both domains, however, is that both academia and industry largely disregard the end user of those translations when assessing translation quality and so, the acceptability of translated and, indeed, source content is not regularly formally measured. Measuring the acceptability (which in this research is defined as usability, quality and satisfaction – see Chapter 3) of translated (and source) text is important because it allows one to identify what impact the translation might have on the end user. One approach that considers the end user is user-centered translation (Suojanen, Koskinen and Tuominen 2015). This approach proposes to gather feedback from end users during the whole translation process and use it to improve the final translation. One of the proposals for TQA in the user-centered approach is the use of usability testing. Usability has also been used for the evaluation of machine translation output in different set-ups, e.g. text comprehension tasks with and without the use of eye-tracking techniques.

The main focus of this chapter is to provide an overview on how TQA is viewed and applied within translation studies and the translation industry in order to identify the different practices for TQA when compared to the user-centered translation model, on which this research is based. The chapter also presents a short section on eye tracking technology in order to describe how it has been used to measure translation quality and how the end user is taken into account with this approach.

2.2 Approaches to TQA in Translation Studies

As previously stated, there is still no agreement on a definition for translation quality and, moreover, "within translation studies, theorists disagree even on how many *categories* of models there are" (Drugan 2013, p.36, emphasis in original).

Several researchers (for example, House 1997; Schäffner 1997; Secară 2005; Fields et al. 2014) believe that evaluation is directly associated with the translation theory being applied to the text and, therefore, "different views of translation lead to different concepts of translational quality, and hence different ways of assessing it" (House 1997, p.1). In order to understand how translation studies has approached the task of assessing translation quality, it is necessary to highlight some of the theoretical approaches to translation evaluation relating to translation as a product. It is important to note that a detailed review of translation theories is beyond the scope of this study, as is a full review of all translation quality models. As Drugan (2013, p.46) notes, "theorists disagree as to how to classify approaches to TQA", and therefore, the review in this section is necessarily limited to the views of translation quality by some of the most influential translation theories, namely: "equivalence", "descriptive" and "functionalist" (skopos) theories.[1]

James Holmes was the first researcher to adopt the term 'translation studies', which for him includes all types of translated texts in all social contexts. The author identifies theories that make a distinction between translation as a process and translation as a product, and emphasises the need for different approaches regarding translation types:

> We need a theory of the translation *process,* that is, the theory of what happens when people decide to translate something. We need a theory of the translation as a *product*, that is to say, what is specific to the translated text; in what ways is it similar to and in what ways is it different from other kinds of texts, literary or other. We need a theory of the translation *function,* that is, how the translation works in the recipient society. And we need a theory of translation *didactics.* (Holmes 1988, p.95, emphasis in original)

For Holmes, translation quality assessment is a part of translation criticism, which represents an improvement from the earlier evaluation practices that were generally considered to be arbitrary and subjective (Lauscher, 2000).

Different theories have different views of what TQA is. The "equivalence" approach defines translation as a reproduction of the source text in the target text, that is, "the attempt to reproduce the source text as closely as possible" (Lauscher

[1] For a comprehensive review on translation theories and quality see Baker 1998; Munday 2008; Pym 2010; Drugan 2013; House 2015.

2000, p.151). Criticism of this approach relates to the fact that “the target text can never be equivalent to the source text on all levels” (ibid.) and, therefore, theorists have differentiated between types of equivalence (Nida 1964; Catford 1965; Baker 1992; Pym 2010). One of the first systematic models for translation quality assessment comes from Reiss (1971), who builds on the concept of equivalence, suggesting specific translation methods according to text types. In her model, “a translation is deemed good if it achieves optimal equivalence” (Lauscher 2000, p.151). Critics of this approach claim that “optimal equivalence” is too vague and that there is no explanation of how to classify text and language functions (see House 2015, p.15).

The “descriptive” approach introduced by Toury (1995) characterises a shift in the equivalence debate. This approach rejects the prescription of the equivalence notion and sees the target text (TT) as the starting point for a translation analysis (Williams 2013). Criticisms of this approach typically point to the lack of emphasis on translators. For House (2015, p.12), the descriptive theory has an overly broad view of what translation is, “which makes it impossible […] to clearly define criteria for translation quality assessment”.

The “functionalist” (or skopos) approach was proposed by Reiss and Vermeer (1984). According to this theory, “it is the purpose of a translation that determines the translation strategy and the shape it takes in the host culture” (Williams 2013, p.53); that is, the purpose is the most important factor in translation. Some of the views in this approach analyse the source text (ST) compared to the TT, and some only focus on the TT. House (2015, p.11) does not consider the functionalist approach useful for TQA since it is not clear “how one can determine whether a given translation fulfils its skopos”. Other criticisms are that the theory is not as applicable to literary texts as it is to more operative/informative texts and that it neglects the ST, and the reproduction of TT features on the micro-level (see Schäffner 1998).

Drugan (2013) states that although translation quality has been a significant focus for translation studies since its recognition as an academic discipline, “few theorists have published detailed, reproducible TQA models for human translation with an indication of the text types on which they were tested” (ibid., p.46) and

have tended instead "to critique other's (sometimes inferred) approaches or tease out what various translation theories seem to imply for TQA" (ibid., p.50). The author lists four specific models for TQA—House's model of TQA (House 1997), Larose's teleological model for translation assessment (Larose 1987), Al-Qinai's empirical, eclectic model for TQA (Al-Qinai 2000) and Williams' argumentation-centred approach to TQA (Williams 2004)—and states that these models are unusual to translation studies because they have been tested and present detailed examples of how translation is assessed. Drugan, however, claims that these models are not fully inclusive of professional approaches and, moreover, suggests that "a new and useful way to classify approaches might be precisely to separate those which are purely academic and those which are designed, adopted and refined based on ongoing applied professional experience" (ibid., p.49).

For Munday (2008), even though there has been a movement away from prescriptive approaches to translation, new perspectives to translation have continued to emerge in recent years, "each seeking to establish a new 'paradigm' in translation studies" (ibid., p.15). The author affirms that translation methodology has evolved and has become more sophisticated but there is still "considerable divergence", as the object of study has changed over time from translation as connected to pedagogy to the "study of what happens in and around translation, translating and now translators" (ibid.). For Munday, this shift on the object of study of translation has allowed for a framework in which the choice of theory and methodology is crucial and "depends on the goals of the research and the researchers" (ibid.).

From the above, it seems reasonable to assume that the definition of translation quality, and the various models that purport to measure it, remains a source of intense disagreement within the field of translation studies. The translation industry faces the same problem; therefore companies have tended to apply their own methods for translation quality evaluation in accordance with their own goals. The next section briefly discusses how the translation industry has approached this issue.

2.3 Industry Approaches to TQA

In the translation industry, quality is largely related to customer opinion (O'Brien 2012; Drugan 2013) and yet, quality evaluation in the translation industry is "managed by gatekeepers in the supply and demand chain who work with static evaluation models […] applying penalties and maintaining thresholds with little, if any, input from customers" (O'Brien 2012, p.55). The rise of machine translation systems in the translation industry has also contributed to making TQA a much debated topic, since "human and machine translation […] quality evaluation methods have been fundamentally different in kind, preventing comparison of the two" (Lommel et al. 2014).

The evaluation models used in the industry are still heavily premised on error-based approaches, where errors are counted and classified in random samples by a translator/linguist/reviewer. In the late 1990s, lists of error types such as the Localization Industry Standards Association (LISA) QA Model[2] began to be used in the industry and, up to this day, many company-specific models for translation quality evaluation have been customised from the LISA model (O'Brien 2012). The LISA QA model consists of a list with types of errors categorised as 'minor', 'major' or 'critical'. Each segment of the translated text is assigned a score depending on the type of error it contains, and then a total score for the whole evaluation task is calculated. A translation receives the status of 'pass' or 'fail' depending on the threshold defined by the evaluator. One of the major limitations of this type of QA model is their "one-size-fits-all" approach, which restricts the comparison of evaluation models (Lommel et al. 2014). One example of this problem is found in O'Brien (2012), where the author distributes a benchmarking exercise among eight big companies in order to identify the types of evaluation models these companies implement in their translation process. O'Brien observes that each company has their own model heavily based on the LISA QA type, in which the companies try to identify and rank errors according to their severity. More interestingly, the author compares each model (also with LISA) and finds that, not only can error categories

[2] There is no public reference available for the LISA QA Model as the organisation is now defunct.

vary widely between models (e.g. 'language' is a common category which may include 'grammar' in one model, but not 'syntax' which, in turn, is a category of 'language' in another model), but the error penalties applied are not comparable between models (what is identified as 'critical' in one model, may be identified as 'major' in another). As Drugan notes, "there is little consensus inside the industry as to definitions of 'major' and 'minor' errors" (2013, p.39), which makes the replication of a TQA model rather difficult.

However, the move to quantify, and thus increasingly standardise, TQA has gained traction in recent times, with the development of an ISO certification of translation parameters, namely, the ISO/TS 11669:2012. The ISO 11669 is a guideline standard that "provides guidance concerning best practices for all phases of a translation project" (ISO/TS 11669:2012) and features a framework for a structured translation specification consisting of 21 translation parameters in five categories: source content, requirements for the target, production tasks, environment, and relationships. Translation quality is defined by the standard as:

> When both requesters and TSPs agree on project specifications, the quality of a translation — from a workflow and final delivery perspective — can be determined by the degree to which the target content adheres to the predetermined specifications (ISO/TS 11669:2012)

For Muegge (2015, p.555), the standard is "a major evolutionary step forward" when compared to previous standards such as the ASTM F2575[3], since it devotes a great space to terminology management. For Muzii (2014, p.424), however, the ISO 11669 is a long list of parameters that builds upon "vague, blurry, and subjective criteria for quality assessment from the archetypal academic scenario" of a traditional "error-catching approach" (ibid.)

This lack of consensus in defining TQA in the industry, even when standards are set, is perhaps to be expected, where translation companies concerned with individual circumstances tend to apply their own working definitions of "quality", given that translation quality in this context can significantly affect important factors (including profitability, loss of clients, etc.).

[3] http://www.astm.org/Standards/F2575.htm [Last accessed: 08 March 2016]

2.4 Filling the Gap

As stated previously, even though there has been increased interest in measuring translation quality, a single standard for TQA still does not exists. Koby et al. (2014) acknowledge the gap between industry and academia and state the need for agreement on an objective way to measure translation quality. The authors themselves disagree on a single definition of translation quality, offering both a broad *and* narrow definition in its place. The broad definition states that:

> 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* [my emphasis] (Koby et al. 2014, p.416)

This broad view of translation classifies several activities such as summarisation, localisation, transcreation (creative translation), and gisting (raw machine translation) as "translation". In this view, specifications relative to the *audience* and purpose should be made explicit whenever possible; that is, requesters and providers should negotiate requirements and discuss the end-users' needs and state them as specifications before the translation process begins. This broad view considers that there are no absolute specifications that can be applied to all projects.

In contrast, the narrow definition categorises translation as text-centric, that is, activities such as summarisation, localisation, etc., are not considered to be "translation":

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

Moreover, the narrow view suggests that explicit specifications are often unnecessary because requesters and end users do not always know what specifications a project requires.

According to the authors, from a quality management perspective, the narrow definition can be seen as a special case of the broad definition, whereas from the point of view of the narrow definition, the broad definition of translation should be viewed as "covering translation quality management rather than just translation quality" (ibid., p.417). Regarding error categories, both views entail that multiple error categories can be used to create a TQ metric, however, the views diverge on the error categories, which are essential to create a TQA metric. Whereas the narrow view holds that meaning transfer, terminology, domain-specific writing quality and domain-independent target–language accuracy are essential categories, the broad definition offers that some error categories are dependent on specific situations, where, for instance, time-sensitive work might not require all of the error categories that the narrow definition determines to be essential (ibid.). Although the authors disagree on a singular definition for translation quality, they agree that a method for measuring translation quality "should emphasize identifying problems that can be corrected" and that "any effort to measure translation quality is doomed to confusion without an explicit definition of translation quality" (Koby et al. 2014, p.416).

The Translation Automation User Society (TAUS) has been at the forefront of attempts to benchmark indicators for effective translation quality assessment. In a recent report, TAUS considers different variables such as communicative function, end user requirements, context, mode of translation (human translation (HT), raw MT output, and post-edited MT), as well as content profiling as precursors to translation quality assessment. The Dynamic Quality Framework (DQF) (O'Brien et al. 2011) outlines that rather than handling problems after the translation process, quality should be considered before the translation process begins.

Another project that aims to standardise TQA is QTLaunchPad, a two-year (2012-2014) EU-funded collaborative project aimed at "identifying quality barriers in translation and language technologies and preparing steps for overcoming them" (Doherty et al. 2013, p.3) by: assembling and providing data and tools for QA;

creating shared quality metrics for both human and machine translation, and improving automatic quality estimation; and expanding existing resource-sharing platforms for MT research (Uszkoreit and Lommel 2013). The QTLaunchPad project feeds into the QT21 machine translation project[4], and aims at developing improved statistical and machine-learning based translation models for challenging language and resource scenarios; improved evaluation and continuous learning from mistakes; and strong focus on scalability, minimising reliance on data.

One of the outcomes of the QTLaunchPad project is the development of the Multidimensional Quality Metrics (MQM) framework, which describes and defines translation quality metrics to assess the quality of translated texts based on the identification of textual features and specific types of issues in the text (Görög 2014). MQM was developed in order to address the shortcomings of the previous quality evaluation models, and even though it takes many principles from the LISA QA Model, MQM was designed to avoid the problems associated with the one-size-fits-all models by "defining a model to declare multiple metrics rather than one single metric" (Lommel et al. 2014, p.459). It was also designed to be flexible when working with other standards, so that the quality of the whole production cycle can be evaluated. The underlying concept is that MQM should comprise a complex model of which users can use just the parts required for some actual purpose (Uszkoreit and Lommel 2013).

The DQF and the MQM are the most recent initiatives that attempt to standardise TQA. In an effort to bridge the gap between the definitions and specifications of these two models, TAUS and DFKI[5], now as part of the QT21 project, have harmonised the metrics and, presently, the TAUS DQF Error Typology is a recognised subset of MQM. According to their website[6], the harmonisation process required substantial modification to both MQM and DQF, but now "users

[4] The project started 1st February 2015 and is expected to end on 31st January 2018. See http://www.qt21.eu/ [Last accessed: 06 May 2016]
[5] http://www.dfki.de/web [Last accessed: 08 March 2016]
[6] For detailed information, see http://www.qt21.eu/mqm-definition/definition-2015-12-30.html [Last accessed: 08 March 2016]

will no longer have to choose between the two [models] because they will share the same underlying structure".[7]

As seen from the discussion above, efforts have been made in order to move to a more dynamic quality model that can take into consideration different views of translation quality, including the view of end users as suggested by the broad definition of translation discussed above. For Koby et al. (2014, p.417), translation studies and the translation industry "need a way to compare different sorts of translation as objectively as possible, with an emphasis on identifying problems" and, the metrics should be "built on a well-defined foundation including at least clearly stated definitions of translation, quality, and translation quality" (ibid., 420). In the next section, the view of user-centered translation (UCT) in which the end user is the primary focus is presented.

2.5 User-Centered Approach

With the introduction of the functionalist approach, the purpose of the translated text has been one of the main focus of translation. Translators are encouraged to consider what the end user needs from the translation and to translate the source text accordingly. However, as outlined in the previous sections, there is a gap between theory and practice in the translation field, and, regarding user-based approaches, the gap between the theory of how to account for the end user and its practice also holds true. As Nord (2014, p.35) outlines, critics have questioned "how translators know what the audience expects of a translation", while also highlighting that "the development of functionalist teaching material is still in its infancy" (ibid.). Suojanen, Koskinen and Tuominen (2015) propose a user-centered approach to translation practice and research that intends to address the problem mentioned by Nord by offering "practical tools and methods for making reader-orientedness an explicit part of the translation process" (Suojanen, Koskinen and Tuominen 2015, p.1).

[7] https://www.taus.net/think-tank/news/press-release/dqf-and-mqm-harmonized-to-create-an-industry-wide-quality-standard [Last accessed: 08 March 2016]

The user-centered translation (UCT) approach can be seen as closely related to functional translation theories as it also focuses on the purpose of the translation. However, for UCT, the users have a central role in the production of the translation and their preferences should be given priority over the client's if there is a clash between the two. User-centered translation examines usability research approaches from the perspective of translation in order to develop a model in which the end user is considered consistently throughout the translation process. The UCT approach is heavily based on the concept of user-centred design (UCD) from usability research, which emphasises the importance of gathering information about the user and identifying ways in which this information is brought into the software development process (ibid., p.4). Considering user-centred design, the UCT approach claims that the involvement of the end user from the start of the translation process is essential because the end user can comment on and/or test the text during all stages of the translation, which can feed back into the translation process:

> …in user-centered translation, information about users is gathered iteratively throughout the process and through different methods, and this information is used to create a usable translation (Suojanen, Koskinen and Tuominen 2015, p.4)

The UCT approach presents several similarities with the broad definition of translation (Koby et al. 2014) seen in the previous section. For example, as in the broad definition of translation, UCT includes machine translation as one activity of "translation" and states that "the decision on whether or not to use machine translation and on how to ensure its usability can be made within the context of the UCT process" (Suojanen, Koskinen and Tuominen 2015, p.6). Also, as the broad definition of translation, the UCT approach defends 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.

As mentioned previously, Suojanen, Koskinen and Tuominen's UCT approach describes concrete tools and methods that the translator can use in taking the end user into account. UCT proposes profiling the future users of translation by means

of mental models, such as: i) intratextual reader position - which are reader positions built into text; ii) audience design - which refers to recipient-oriented shaping of translation based on five categories: addressees, auditors, overhearers, eavesdroppers and referees; and iii) personas – which are imaginary characters representative of real groups of users. Although the UCT's position for offering the model is academic, the authors believe that theory and practice should not be separated and, therefore, the model is also claimed to be a framework for translation practice that can be used by practising translators and also used in the translation industry. The authors argue that due to the competitive market situation, companies have to become more flexible and innovative and the user-centered translation approach allows translation companies to create "new value for customers" and redefine "the products and services offered" (ibid., p.2).

Regarding translation evaluation, the authors argue that traditional TQA practices suffer from "end-of-the-line" problems, that is, TQA mostly "focuses on measuring the end product" in which "any changes can be costly both financially and in terms of missed deadlines" (ibid., p.128). They add that a consensus on a definition of quality is rather difficult to achieve since views of error-based approaches, in which lists of "criteria for a successful translation" (ibid.) are used, are rather subjective. Moreover, Suojanen, Koskinen and Tuominen affirm that UCT can be an alternative basis for evaluation as it "concentrates on imagining what kind of a process will produce a variety of successful translations to serve the needs of different commissions" (ibid.). The authors propose usability heuristics for user-centered translation, and although those heuristics can be used either for assessing the translated texts (product) or for generating the texts (process), the authors emphasise the text generation side, since translation studies has several source-text analysis models whereas "translators need more concrete tools to be able to produce a target text appropriate for its users" (ibid., p.89). Their proposed heuristics framework include: 1) match between translation and specifications; 2) match between translation and users; 3) match between translation and real world; 4) match between translation and genre; 5) consistency; 6) legibility and readability; 7) cognitive load and efficiency; 8) satisfaction; 9) match between source and target

texts; 10) error prevention. The authors claim that this is a generalised list and translators can use it to develop their own contextualised versions.

Another claim for the use of UCT heuristics is that traditional TQA practices may be intimidating for translators as they feel that it is a negative criticism of translators' performance, whereby assessment is focused on random segments instead of focusing on crucial parts of the text. In comparison to traditional TQA, the authors claim that:

> UCT is an interactive process, and usability assessments are completed incrementally, to verify translation strategies and textual choices before the text is finalized. Errors, especially translation mistakes in comparison to the source text, 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, Koskinen and Tuominen 2015, p.129)

The authors admit that the UCT model has received criticism regarding the feasibility of introducing usability tests with actual users into a traditional translation project in the translation industry, and acknowledge that while mental models can be easily introduced into a translation project, usability testing may not be so easy, and in addition, it could be very costly.

For this reason, Suokas et al. (2015) present an experiment aimed at applying the UCT usability tests in an actual translation project involving web-based course material designed for international students. The experiment consists of two tests: the first test provides a reading activity designed to collect the users' subjective comments and to identify potential usability issues on a textual level. The second test is a task-based model whereby participants perform a task designed to collect usability issues while participants use the text. The selected content is the web-course material translated by translation students into English and consists of two excerpts: the first is an introduction to translation memory software at a general level, and the second an instructive text on how to start a new project with the translation memory software used on the web-course.

In the first experiment, the content used was a first draft of the translations, and fourteen participants from different nationalities and languages were asked to read the text from the perspective of a student of the course. They were asked to

take down notes of parts that caught their attention, considering their own perspective as a student. After this activity, a focus group took place. Questions about the grammar and style (unfamiliar vocabulary, misspellings, long sentences, etc.) were asked and a follow-up on the participants' comments was also introduced. The participants noted that the text contained some issues concerning sentence structure and wording, long sentences, formality and register. These comments were analysed and used to improve the usability of the translation of the course material.

In the second experiment, four participants were not present and, therefore, the experiment was carried out with ten participants. The session was held a week after the first experiment and the translated instructions that were improved via the previous participants' comments were used. The participants were asked to create a new project in the translation memory system using the instructions with a time constraint of thirty minutes. Similar to the first experiment, a focus group was held after the experiment was finished in order to identify participants' perceptions on how helpful the instructions were. Regarding goal completion, eight participants successfully created a translation memory project within the time limit, whereas the remaining two participants partially completed the task. Regarding the focus groups, the researchers noted fewer comments from the participants, which the authors considered an indication that the participants did not encounter many difficulties when using the instructions. The authors also report that the participants agreed that the text seemed to fulfil its purpose.

In summary, Suokas et al. claim that the UCT usability testing is an efficient approach since it enables the translators to gather information during the first phase of the translation process, which, in turn, allowed them to improve the usability of the final translations. The authors add that when compared to traditional translation quality assessment, which is generally performed by translators or reviewers, usability testing allows the intended target audience to be taken into consideration. They conclude that the usability test applied in the translation project provided positive results and that further experimentation should be supported.

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 method approach that aims to analyse and evaluate different types of translations (namely machine translation, post-editing, and a small amount of human translation) across three languages and the English source text.

In order to contextualise the present research, an overview of machine translation evaluation approaches are presented in the next sections, including human and automatic evaluation, as well as post-editing and usability evaluations. This will be followed by an overview of the use of eye tracking in cognitive research and how it has been deployed for translation evaluation.

2.6 Machine Translation Evaluation

Machine translation is the process of automatically translating text from one natural language into another (Dorr et al. 1999); therefore, machine translation evaluation (MTE) is the task of assessing the quality produced by an MT engine. MTE is a long-standing practice that has been under study since the early years of machine translation itself. The lack of agreement within the field of translation studies and within the translation industry on what is a 'good' translation has also led to many approaches for MT evaluation. Therefore, translation quality can be considered from a range of different perspectives and there is no single design that suffices to address all evaluation purposes (Hovy, King and Popescu-Belis 2002). However, MT evaluation has not yet given much consideration to the usability of MT output, having the end user as the evaluator.

There are several reasons why one would be interested in MTE (ibid.): In academia, MTE is the means to determine whether or not new methodologies have led to quality improvements. Ultimately, evaluation helps drive research and development. In industry, evaluation is necessary in order to provide evidence of quality in order to sell MT commercial solutions. Comparatively assessing translation quality may help end users of MT to decide which system they want to use. There are also several aspects that can be considered when evaluating the

performance of MT systems, the major ones being i) translation quality, ii) system performance, and iii) cost and usefulness (Dorr et al. 1999).

MT quality can be assessed both manually and automatically. On the one hand, automatic evaluation has commonly been accepted as being objective and cheap, however it has been claimed that it is less comprehensive than manual evaluation and does not indicate the type of quality problems the translated text contains (Uszkoreit and Lommel, 2013). On the other hand, manual evaluation is often claimed to be subjective and can be expensive to perform (Callison-Burch et al. 2011; Bojar et al. 2011); however, manual setups can assess complex linguistic phenomena, such as error types, adequacy and fluency (Section 2.6.2).

Several researchers have addressed the problem of measuring translation quality for MT. The Defense Advanced Research Projects Agency (DARPA) MT evaluation constitutes one of the earliest efforts in MTE in the 1990s. The DARPA Initiative lasted four years and aimed at developing new MT approaches as well as methodologies for evaluating MT systems (White and O'Connell, 1996) (see Section 2.6.2). Another effort aimed at tackling MTE issues is the Framework for the Evaluation of MT (FEMTI) that aims at drawing an overall perspective of all the MT evaluation metrics according to the evaluation purpose, the main goal of which is to "build a coherent picture of the various features and metrics that have been used in the past, to offer a common descriptive framework and vocabulary, and to unify the process of evaluation design" (Hovy, King and Popescu-Belis 2002, p.44). The Workshop on Statistical Machine Translation (WMT), held annually since 2006, is a very well-known venue for research in the MT field. The workshop features a shared translation task for evaluating MT systems, which is claimed to act as an extensive manual and automatic evaluation of machine translation performance (Callison-Burch et al. 2007).

In the following sections, an overview of the state of the art for MT evaluation is presented, starting with the automatic metrics (2.6.1); human judgments for MT evaluation (2.6.2); post-editing as MT evaluation (2.6.3); and finally, usability methodologies applied for MT evaluation (2.6.4).

2.6.1 Automatic metrics

Automatic evaluation for MT is a very active area of current research. Many researchers are dedicated to evaluating and improving these automatic metrics as well as proposing new ones (Papineni et al. 2002; Snover et al. 2006; Lavie and Agarwal 2007). The main purpose of automatic evaluation metrics is to compare the output of an MT system to one or several reference translations, which are generally claimed to be gold standard human translations. Automatic metrics, therefore, try to measure how close the MT output is to the reference translation (Koehn, 2010).

The first automatic metrics used in MTE came from the speech field, e.g. WER used by Nießen et al. (2000). Subsequently, BLEU (Papineni et al. 2002) was proposed and by showing correlation with human judgement, it became the official metric of the MTE series from the National Institute of Standards and Technology (NIST). Other common automatic metrics include NIST (Doddington, 2002), METEOR (Lavie and Agarwal, 2007) and TER (Snover et al. 2006). Although several new metrics have been developed since then, BLEU is the standard metric for most research papers together with METEOR and TER. The Workshop on Statistical Machine Translation compares different metrics to measure the correlation with human judgements and has found that several new generation metrics outperform BLEU regarding correlation with human judgements (Callison-Burch et al. 2009).

One of the main arguments for using automatic metrics is that they have the advantage of requiring minimal human labour. As opposed to human assessments, they do not require bilingual speakers to assess the translation, which makes the assessment cost-effective. However, it is important to note that translators are needed in order to create the reference translation used in the process. Additionally, gold standard quality of the human reference is assumed, but often not verified.

Automatic MTE also provides rapid feedback and is often used on an on-going basis during system development to test changes in the system. The problem with automatic metrics is that their ability to assess syntactic and semantic equivalences in MT outputs is yet to be proven, since they lack linguistic analysis and understanding, and face just as many challenges as MT itself. Although METEOR

allows for non-exact matches such as synonyms and paraphrases, more complex syntactic and semantic equivalences are far from being recognised. To overcome that deficiency, a number of automatic metrics with deeper linguistic analysis have been proposed (Giménez and Màrquez 2008; Padó et al. 2009; Liu, Dahlmeier and Ng 2010).

Recently, attempts have been made in order to propose metrics to evaluate machine translation at the document level, which could lead to a more precise way to measure the coherence of an automatically translated text. These include Giménez et al. (2010) who propose to use co-reference and discourse relations in MT metrics; Wong and Kit (2012) who apply lexical cohesion to existing sentence-level evaluation metrics; Guzmán et al. (2014) who incorporate discourse structures to complement existing MT evaluation metrics.

In summary, automatic metrics evaluate MT quality by comparing the MT output with a reference translation and are often used on an on-going basis by developers.

2.6.2 Human Evaluation

Human judgements of machine translation quality have been used since the first experiments with MT. Human evaluation consists of having human participants, either monolingual or bilingual, judging the output of an MT system according to several different features. The type of evaluation depends on what is intended to be measured; therefore, the profile of the participants may also vary.

Some of the most frequently used manual metrics are the ratings of fluency and adequacy. Fluency and adequacy are generally measured via a Likert scale, where the evaluator is asked to assign a score to the translated segment. In White and O'Connell (1994, p.136) fluency evaluation assesses "intuitive native speaker senses about the well-formedness of the English output on a sentence by sentence basis", while adequacy, compared against expert translations, measures the extent to which the meaning of the reference translations is present in the MT output. To evaluate fluency, the evaluator needs to be a fluent speaker of the target language. There is no need for the evaluator to know the source language, since fluency does

not require the automated translated sentence to be an accurate translation of the source. To judge adequacy, however, the annotator must be bilingual in both the source and target language in order to judge whether the information is preserved across translation, although, in some adequacy evaluation setups where the source is compared against high quality human translations of the source sentence, the annotator could be fluent only in the target language.

Error analysis is another common practice for evaluating MT output. It consists of the identification and classification of individual errors found in the MT output: it is "a means to assess machine translation output in qualitative terms, which can be used as a basis for the generation of error profiles for different systems" (Stymne and Ahrenberg 2012, p.1785). This type of evaluation allows for the identification of particular strengths and problem areas of MT systems and to diagnose what went wrong and which research direction to take (Flanagan 1994; Correa 2003; Vilar et al. 2006; Llitjós 2005; Stymne et al. 2012). Error analysis has also been used to identify problematic passages in the MT which can be fixed after the post-editing process, as well as passages that remain problematic even after PE is implemented (Daems, Macken and Vandepitte 2014). This approach provides rich data which can also be used to improve post-editor training.

Other frequent methods used to assess MT quality through human judgement include ranking translation, which consists of ranking translated sentences by an MT system from best to worst (Callison-Burch et al. 2007) or, in some cases, the participants are asked to assign scores to each translated sentence/segment on a pre-determined scale (LDC 2002). Reading comprehension or even comprehension tasks using the system output (Fuji 1999; Jones et al. 2005) is also one of the methods (see Section 2.6.4 – usability evaluation for detailed description). It is important to note, however, that reading comprehension tasks are rather rare in the MT evaluation field. Additionally, measuring the amount of work required to post-edit the system output (see Section 2.2.3), such as time (Sousa, Aziz and Specia 2011) and keystrokes has also been explored.

As mentioned previously, one of the first major projects that aimed at defining human evaluation metrics[8] was DARPA's project on MTE (White, O'Connell and O'Mara 1994). Evaluators were asked to assess automatically translated sentences according to the concepts of fluency and adequacy, assigning a score from 1 to 5. Adequacy assessment was performed by comparing the MT output against the source text (White, O'Connell and Carlson 1993) and against professionally-produced human reference translations. A reading comprehension task was also part of the MTE methodology where the evaluators were asked to answer questions about the text.

The Workshop on Machine Translation adopted human judgements as a primary methodology for assessing translation quality in its 2007 edition (Callison-Burch et al. 2007 and 2008), while the first two years of the workshop were focused on automatic metrics. The evaluation process was based on the concepts of fluency and adequacy (on a scale from 1-5) and the methodology was premised on ranking translations relative to each other. In 2009, Callison-Burch et al. introduced the evaluation of post-edited sentences. The authors do not clarify whether there were qualified translators involved in the post-editing process. The annotators were asked to post-edit the sentence to be "as fluent as possible" without seeing the reference. Following this, they were asked to judge the post-edited translations by annotating with yes/no, whether the sentences were fluent considering the reference sentence.

Recently, crowdsourcing has become popular in the field of translation (crowdsourcing translation)[9] and it has also been applied for human evaluation of machine translation. Callison-Burch (2009) proposes several ways to evaluate MT output by making use of Amazon's Mechanical Turk[10], a platform to crowdsource content that is based on tasks. The author experiments with crowdsourcing for ranking translation from best to worst; creating multiple references by translating the source text; detecting machine translated sentences by selecting the sentences

[8] Although The ALPAC (Automatic Language Processing Advisory Committee) report had already used human ratings of intelligibility back in 1966.

[9] Crowdsourcing translation has often been used as synonymous for community translation, user-generated translation and collaborative translation (O'Hagan 2011).

[10] https://www.mturk.com/ [Last accessed 14 March 2016].

that 'look like' they came from an MT system; post-editing of machine translation; judging post-edited translation by ranking those that are close to the reference translation; reading comprehension tests by i) reading the text and creating questions about it and, ii) reading the text and answering questions about it.

Crowdsourcing has also been explored in discussion forum contexts (Mitchell 2015; Mitchell, O'Brien and Roturier 2014). In Mitchell, O'Brien and Roturier (2014), the authors report three quality evaluation methods for community post-edited content. First, the authors ask for community members from the German Norton Community[11] to post-edit twelve texts taken from the English-speaking community. Afterwards, the post-edited content was evaluated for fluency and fidelity by domain specialists, an error annotation of MT was performed by a trained linguist and fluency was rated by community members. The results show that the community evaluation and the evaluation performed by domain specialists have similar results.

Crowd assessments of MT may allow evaluations on a large scale while being cost-effective, however, as Zaidan and Callison-Burch (2011, p.1221) outline, "soliciting translations from anonymous non-professionals carries a significant risk of poor translation quality". Even though there may be professional translators in crowd communities (O'Hagan 2011), the quality of work in crowdsourcing is generally not guaranteed, since the crowd may employ as little time as possible in the tasks or even employ someone else to do the test for them (Graham et al. 2013). In order to tackle these problems, several methodologies have been developed to filter the evaluations. One method is to compare crowdsourcing with expert evaluations (Callison-Burch 2009; Goto, Lin and Ishida 2014); however, it is not clear what some authors mean by "expert", i.e., some may refer to trained translators, whereas others may refer to computational linguists who develop machine translation systems, as is the case of Callison-Burch's methodology (2009). Another methodology presented to tackle the crowdsourcing problem is that of Graham et al. (2015) who propose to collect judgements on a continuous rating scale, whereby the crowd develops their own individual assessment strategy by assessing each translation in isolation (i.e. not comparing different translations at

[11] http://www.de.community.norton.com/ [Last accessed 14 March 2016].

the same time as is done, for example, in ranking translation tasks). According to the authors, this methodology has the advantage that agreement with the expert is no longer required, and more meaningful statistics can be computed.

As seen from the above, there are several ways to apply human evaluation for MT. The main advantage of using human evaluation for MT is that it can assess deep linguistic information that provides reliable insight for error analysis, which in turn, can help to understand the actual linguistic strengths and weaknesses of an MT system (Gaspari et al. 2014). However, human evaluation can be very subjective, suffering from disagreement when annotators are not well trained for the task (Callison-Burch et al. 2011; Bojar et al. 2011); it may also be time consuming (depending on the scale of the task) and expensive.

2.6.3 Post-editing in Machine Translation Evaluation

Post-editing (PE) is the practice of modifying pre-translated text so that a quality need is met. There are generally two distinct degrees of post-editing: so-called '*light* post-editing' and '*full* post-editing'. Light post-editing has a quick turn-around and only essential errors are corrected, whereas full post-editing requires more corrections for a higher quality, with a slower turn-around (O'Brien, Roturier and De Almeida 2009).

Krings' (2001) seminal work on the post-editing process divided post-editing effort into three categories: *temporal*, *technical*, and *cognitive*. Many works have built upon Krings' work and, in recent years, the task and process of post-editing has received significant attention in (e.g. De Almeida and O'Brien 2010; Depraetere 2010; Plitt and Masselot 2010; Sousa, Aziz and Specia 2011; Specia 2011; Koponen 2012; O'Brien et al. 2012; O'Brien et al. 2013; O'Brien et al. 2014; Lacruz and Shreve, 2014; Guerberof 2014; Moorkens et al. 2015; Daems et al. 2015; Carl, Gutermuth and Hansen-Schirra 2015), as MT systems gain space in the industry. Although it is possible to say that there have been great advances in MT, post-editing pre-translated text is still the traditional means for achieving publication quality.

An important use of post-editing is the collection of information that can be used for measuring machine translation quality and diagnosing translation problems. As a result, several tools have been developed in order to capture such information, such as PET (Aziz, Sousa and Specia 2012), Translog-II (Carl, 2012), CASMACAT (Alabau et al. 2013), and iOmegaT (Moran, Lewis and Saam 2014). In the industry, one of the major concerns surrounding translation services is how to quantify the amount of effort that is necessary for post-editing pre-translated text; the purpose of which is to determine whether post-editing MT output would be time and cost-effective when compared to translating the text from scratch. Objectively assessing the post-editing effort has become indispensable, given that it enables companies to optimise their translation process.

Several approaches have been attempted to understand the level of cognitive effort in post-editing pre-translated text, while clarifying what the effort indicators are and what they can be used for. Snover et al. (2006) have measured PE effort in terms of an edit distance, that is, the amount of edit operations (e.g. insertion, deletion, substitution, shift, etc.) that transforms the MT into its post-edited version. Tatsumi (2009) and O'Brien (2011), attempt to determine if automatic metrics correlate with human judgements. Results from both studies suggest that even though there is some correlation between PE effort and automatic metrics, it is not a linear one. Sousa et al. (2011) compare the time spent on post-editing to i) subjective assessments on effort and quality, and ii) automatic metrics of MT evaluation such as BLEU, METEOR and HTER. Results show that sentences requiring less time to be post-edited are more often tagged by humans as demanding low effort. In addition, the PE time has shown positive correlation to BLEU, METEOR and HTER metrics, that is, sentences that required less time to be post-edited scored better for those metrics. Specia (2011) uses post-editing effort classified in terms of time, subjective scores and PE distance to predict the quality of an MT system. Results show that using those effort indicators to train the confidence estimation models produces rankings of translations that reliably reflect their post-editing effort. Daems et al. (2015) examine the impact of different types of machine translation errors on post-editing effort indicators from English into Dutch. Their results show that average MT error weight is a good predictor of six different post-

editing effort indicators such as average number of production units, average duration per word, average fixation duration, average number of fixations, average pause ratio, and pause ratio.

As seen from the above, recent advances in MT have enabled post-editing to become a more common practice in the translation industry, which has led to much research on post-editing effort (Snover et al. 2006; O'Brien 2011; Sousa, Aziz and Specia 2011; Moorkens et al. 2015). However, MT evaluation has not yet considered the *usability* of MT output with the end user as the evaluator. The next section discusses attempts to evaluate machine translation quality from the end-user perspective.

2.6.4 Usability Evaluation of Machine Translation

Despite the considerable focus on MT quality evaluation, the impact of MT on the end user has been significantly under-researched. In fact, apart from a few studies, very little research has been carried out on the impact of different modes of translation (e.g. human translation (HT), raw MT output, light post-editing of MT, full post-editing of MT) on the end user.

Usability methodologies have been known to address the end-user's needs and, as already mentioned, it has been defined as by ISO (2002) "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 2002). This definition of usability is the one this research is based on and, therefore, this section contrasts related work on usability for MT to ISO's definition (the definition of usability will be comprehensively discussed in Chapter 3, Section 3.1.1).

Tomita et al. (1993) compare different MT systems by using reading comprehension tests. The content for reading and comprehension was extracted from an English proficiency exam and then translated into Japanese via three commercial MT systems as well as through the process of human translation. Sixty native speakers of Japanese were asked to read the text and answer the questions. The authors show that reading comprehension is a valid evaluation methodology for MT; however, their experiment does not take into consideration any other

aspects of usability apart from informativeness (measure by the number of correct answers for the comprehension questions).

Fuji (1999), also presents a usability method for evaluating post-edited MT outputs. He proposes reading comprehension tasks in order to measure informativeness as in Tomita et al. (1993) and, moreover, the author adds comprehensiveness and fluency to the evaluation measures. The content used comprises several texts from official examinations of English language designed for Japanese students. Participants are asked to read the text, answer the comprehension questions and judge how comprehensible and how fluent the text is, on a 4-point scale. Following on from this, Fuji et al. (2001) examined the "usefulness" of machine-translated text from two commercial MT systems compared to the English version. The experiment consisted of participants reading the texts and answering comprehension questions. Afterwards, they were asked to evaluate the MT outputs on a 5-point scale regarding comprehensibility and awkwardness. The authors claim their evaluation approach delivered statistically significant results easily understood by the general public.

Guessability and learnability is evaluated by Gaspari (2004) for five online MT systems. The author defines guessability as the "effort required on the part of the user to successfully perform and conclude an on-line task for the first time" (p.75), and learnability as "the effort and time required on the part of the users to familiarize themselves with the satisfactory operation of a web-based application after they have used it already at least once" (p.76). The results of this small-scale evaluation show that different approaches to interaction design can dramatically affect the level of user satisfaction. However, the author focuses only on the user interaction with the MT system and not with the actual output.

Jones et al. (2005) present a usability test where 84 English native speakers answer questions from a machine translated and human translated version of the Defense Language Proficiency Test for Arabic language. Task time and subjective rating were also used to measure usability. Their results suggest that MT may enable an Interagency Language Roundtable Level 2 (ILR) performance (limited working proficiency) but it is not suitable for ILR 3 (general professional proficiency). A shortcoming of this approach is that the subjects were allowed eight hours to

complete the test, and, even though they were requested to take breaks "as needed", the likelihood that some of the participants were too tired to answer all the questions with the same interest is high. Moreover, it is not clear how the authors collected the reading times in the task, since the participants were allowed breaks during the experiments.

Usefulness, comprehensibility, and acceptability of MT technical documents are examined by Roturier (2006). The author claims that a text is deemed useful when the readers are able to solve their problem with the help of the translation. He defines comprehensibility as how well the reader can understand the translation and, for him, acceptability refers to the relevance a text has for its receiver and how textual characteristics are accepted, tolerated or rejected by the receivers. The study uses a customer satisfaction questionnaire to determine whether controlled English rules can have a significant impact from a Web user's perspective. The main drawback of Roturier's approach is that there is no task being performed by the end user as the methodology consists of an online questionnaire.

As can be seen, some limited efforts have been made to assess the usability of machine translation. The main shortcomings of the approaches presented are that they tend not to address all the aspects of usability; therefore, a more comprehensive study on usability and user satisfaction is necessary in order to determine end-users' levels of tolerance with regard to machine translation.

Section 2.7 will discuss the use of eye-tracking methodologies to measure translation usability and end-user evaluation.

2.7 Eye Tracking in Cognitive Research

Eye tracking is the process of measuring an individual's eye movements when interacting with texts and images on a screen, or with objects in the surrounding environment. The number of researchers who use eye-trackers has grown considerably in the past 20 years (Holmqvist et al. 2011), and the technique has now been applied to several fields, such as cognitive science, psychology, human-

computer interaction (HCI), marketing research and translation research, with a few training textbooks to aid the researcher (Duchowski 2007; Holmqvist et al. 2011).

According to Holmqvist et al. (2011, p.10), the earliest eye-trackers date from the late 1800s but were "technically difficult to build, mostly mechanical and not very comfortable for the participants". However, with the improvement of eye tracking technology in the 1970s, the use of the technique has increased, making it possible to link eye-tracking data to cognitive processes through the field of psychological theory (Jacob and Karn 2003).

In present days, eye-tracking measures have become well established as indicators of cognitive effort (Rayner 1998; Radach, Kennedy and Rayner 2004). Research in reading has benefited from the study of eye movements "because they are an inherent behavioural manifestation of the reading process in action" (Radach, Kennedy and Rayner 2004, p.1). According to Rayner and Juhasz (2004), it has been argued that there are two extremes regarding research using eye movements: those who are interested in "eye movements *per se* or questions that are related to perceptual processing during reading" and those who are interested in "using eye movements as a tool to study some aspect of the reading process" (Rayner and Juhasz 2004, p.346) and that both extremes should pay attention to the findings from each other. In this study, our focus is on the latter.

Some of the most common eye-tracking metrics in reading research are fixations and saccades, movements that happen during reading:

> During reading, we move our eyes in a sequence of very fast, relatively well coordinated, movements known as saccades. These movements are interrupted by fixations, periods of relative stability in the position of the visual axis, during which visual information can be extracted (Radach and Kennedy 2004, p.1)

Some of the findings regarding fixations and saccades in reading research state that, when reading, eye fixations last about 200-250 milliseconds and the mean saccade size is 7-9 letter spaces (Rayner 1998; Starr and Rayner 2001). When reading conceptually difficult texts, fixation duration increases, saccade length decreases, and the frequency of regressions increases (Rayner 1998).

Studies on eye movements in reading include the effects of word familiarity on word recognition and text comprehension during silent reading (Williams and

Morris 2004), the effects of word length and complexity on inspection durations (Kliegl et al. 2004), eye movement behaviours for different types of reading tasks (Jakobsen and Jensen 2008), the effect of controlled language rules on readability (O'Brien 2010).[12]

Another field benefiting from eye tracking technology is human-computer interaction. By tracking user's eye movements, researchers are able to identify factors that may impact on the usability of a system interface or a web page, for example. Usability measures the performance of a human subject at a task level by taking into consideration measures such as time to complete a task, percentage of participants succeeding, number of errors, etc. (Karn et al. 1999). Eye tracking technology supplements these measures of usability by allowing the researcher to objectively analyse the amount of time the participants spend looking at an area of interest, that is, a pre-defined region of the screen, the sequence of their eye movements when looking at the UI/webpage, changes in the pupil size (Ellis et al. 1998), as well as the number of shifts of attention. Usability data along with eye tracking measures can help to improve the design of interfaces as well as websites. One example of this is the discovery of the F-shaped pattern (Nielsen 2006). In an F-shaped pattern, the readers first read in a horizontal movement across the top part of the content area; in the second read, they move down the page a little and read across a second horizontal movement that is typically shorter than the previous; finally, in the third read, they scan a vertical line down the left side of the text looking for keywords or points of interest in the paragraph's initial sentences. When the reader finds something they like, they begin reading normally, forming horizontal lines. The discovery of the F-shaped pattern has driven websites to use the F-shaped design and has also driven more research towards web usability and quality, such as searching for the F-shaped pattern while searching versus browsing (Shrestha and Lenz 2007), the study of user's attention (Alt et al. 2012), and user behaviour in web searches for large and small screens (Kim et al. 2015). Other contributions of eye tracking to HCI include the study of advanced design interfaces

[12] For a comprehensive review, see Rayner 1998, Radach et al. 2004.

(Jacob, 1995) and efficacy of information search strategies on menu-based user interfaces (Byrne et al. 1999).

Translation research has also adopted eye-tracking technology as a technique in recent years. This area benefits from the establishment of eye-tracking metrics in reading set by the reading researchers and, although the two fields share similar challenges, translation process research has specific methodological challenges (O'Brien 2009). For Göpferich et al. (2008, p.2) the area still lacks information on "eye-movement behaviour during continuous reading or reading with different purposes in mind, e.g. reading in order to translate". In order to address these and other challenges, several works have been published (e.g. Göpferich et al. 2008; Göpferich, Alves and Mees 2010; O'Brien 2011).

O'Brien (2006) is one of the first studies to use eye tracking as a methodology applied to translation. In her study, O'Brien investigates whether eye-tracking is a useful research methodology for studying translators' interaction with translation memory (TM), and whether eye-tracking measures provide indications of differences in cognitive effort when translators deal with different TM match types. Four professional translators translated a text using a TM tool and then commented on their translation process in retrospective protocols. The results suggest that exact matches require less cognitive effort from translators and that relatively good MT matches require effort similar to high fuzzy matches. Moreover, the results from the study prove eye tracking to be an effective method of research in translation processes. Following this, O'Brien (2008) investigates in more detail the relationship between fuzzy matches and cognitive effort by looking at processing speed metrics (words per second) and pupil dilation. The results show that, when considering processing speed alone, decreasing fuzzy matches means increasing effort. However, when considering the pupil dilation, no linear relationship is detected.

According to Alves, Gonçalves and Szpak (2012), research on the translation process has tried to establish cases of demanding processing in translation, and studies in the area have shown that "eye fixations differ in areas of interest (AOIs) found in source and/or target texts and, thus, suggest interesting implications in terms of reading/writing for translation" (ibid., p.6). Studies in the translation

process area are considerable: Jakobsen and Jensen (2008) analyse differences in fixation duration when reading for understanding, for translating, for sight translation and for written translation. The results suggest that translators assign more cognitive effort to the target text rather than to the source texts, suggesting that target text requires more cognitive effort than the source text. Other works worth mentioning are Jensen's (2009) study which provides a discussion about the relevance of readability indices in measuring text complexity; and Alves et al. (2010), who use annotated corpora in order to identify translation units associated with high levels of cognitive effort during the translation process. Hvelplund (2011) investigates the differences in cognitive effort levels between professional and novice translators during the translation process. The results show that novice translators require more cognitive effort when translating from source to target texts, as well as when switching between different types of cognitive processes. Carl and Dragsted (2012) investigate differences between copying and translation tasks. Results suggest that more processing effort is required during translation than during copying tasks. Also, during copying tasks, translators tend to present more parallel reading and writing activities, whereas during translation tasks, translators tend to resort to sequential reading and writing patterns triggered through target text production problems.

Regarding machine translation, a few studies have attempted to use eye tracking in order to evaluate pre-translated texts. Doherty and O'Brien (2009) is one of the first studies to use an eye-tracker tool to record the fixations and gaze time of translators while reading sentences from an MT system output. In total, fifty French sentences that had been previously judged as 'good' and 'bad' (fifty-five for 'good' and twenty-five 'bad') by human evaluators were presented. Ten native speakers of French were instructed to read the sentences for comprehension and, afterwards, they were asked to record a retrospective protocol while watching the recording of their gaze data. Their results show that gaze time was significantly higher for the 'bad' sentences; however, fixation count was not significantly different. Evaluation comments during the retrospective protocol correlate well with previous human judgements of quality. Building on this, Doherty, O'Brien and Carl (2010) add comparisons between BLEU scores and the eye gaze data gathered

from the previous experiment. They found that gaze time and fixation count correlate well with BLEU scores. However, no significant correlation was found with pupil dilation and fixation duration. Moreover, the study found that eye tracking measures correlate well with human evaluation of MT and it is an effective methodology for MT evaluation.

Another study using eye tracking for MT is that of O'Brien (2011) who investigates correlations between two MT automatic metrics (TER and GTM) and post-editing productivity. Post-editing productivity is measured via processing speed (number of words post-edited per second) and cognitive effort (fixation count and fixation duration). Seven translators were asked to post-edit 60 segments with no time constraint imposed while having their eyes tracked. The results show that processing speed, average fixation duration and fixation count per word correlate well with the GTM and TER bands of scores and, moreover, the author concludes that there are reasonable correlations between the two metrics and actual post-editing.

Predictors of cognitive effort in machine translation post-editing are investigated by Vieira (2014), who examines source-text and machine-output features as well as translator's individuality (such as working memory capacity and source language proficiency) in order to determine their impact on cognitive effort when a post-editing task is implemented. Thirteen native speakers of English with different profiles (education background, translator experience) were asked to lightly post-edit two texts (divided into 6 passages) with no strict time pressure, and then rate perceived cognitive effort on a 9-point scale. Cognitive effort was measured via average fixation duration and fixation counts. Results show that the MT automatic evaluation metric METEOR is significantly correlated with all measures of cognitive effort, and post-editing was perceived as more effortful in sentences with higher ratio of fixations by translators with low proficiency in the source language.

More recently, Carl, Gutermuth and Hansen-Schirra (2015) present an empirical comparison of three translation tasks using eye tracking and key logging. Twenty-four German native speakers (twelve professional translators and twelve students) were asked to perform translation from scratch, light post-editing of the

MT output of Google Translate, and monolingual light post-editing (editing) of six English source texts. After the tasks, the participants were asked to answer a questionnaire judging their own performance and the quality of the MT output according to grammaticality, style and accuracy. The results of the study show that in general, 65% of the participants were 'somewhat' or 'highly' satisfied with their performance in the post-editing task, also when compared to the editing tasks. However, 83% of the participants answered that they preferred the task of translation from scratch than the post-editing task and, moreover, 78% said they would rather have translated the texts from scratch than post-edited. Furthermore, 69% of the participants answered that they had to post-edit 75-100% of the MT output. The authors found that results from the questionnaire, however, do not correlate with the gaze data collected from the participants as the fixation duration and fixation counts show that source text complexity seem to influence the cognitive effort during translation tasks but not during the post-editing task. Moreover, their inefficiency metric[13] show that post-editing effort is not evenly distributed over the text which contradicts the translators' statement that they had to post edit between 75 to 100% of the text.

As can be seen, there is a growing use of eye tracking methodology in translation process research but few attempts have been made to specifically assess MT quality and post-editing effort with the technique. Furthermore, very few studies have considered the usability of MT texts from the end-user's perspective. In the next section, an overview of how eye-tracking methodology has been implemented in order to measure usability of machine translation is presented.

2.7.1 Eye tracking in Usability Evaluation of Machine Translation

Measuring the usability of translated texts is crucial for a better understanding of how end users engage with translations and the type of errors that may have an impact on the end users. As previously discussed in Section 2.6.4, very little research has been carried out on the impact of different modes of

[13] Inefficiency score is defined as: InEff = insertions + deletions / length of final translation – where a high InEff value indicates a larger amount of editing activity (less efficiency).

translation on the end user, while the studies available (e.g. Tomita et al. 1993; Fuji 1999; Fuji et al. 2001; Jones et al. 2005; Roturier 2006) have not taken all the aspects of usability into consideration. This section discusses the studies conducted using eye tracking in order to measure usability of different translation modes.

Doherty and O'Brien (2012) is the first study to use eye-tracking techniques to measure the usability of machine-translated texts via the end user. They conduct a study to compare the usability of raw machine translated output for four target languages (Spanish, French, German and Japanese) against the usability of the source content (English). In this study, usability is defined by effectiveness, efficiency, and satisfaction (ISO 2002), where effectiveness is measured via goal completion (number of successful tasks) and efficiency is the number of successful tasks divided by the total task time. The English documentation for a well-known online file storage and sharing service was selected and modified to produce six sequential tasks. Afterwards, the texts were translated via a freely available MT system in order to create a strong ecologically valid scenario, where users looking for online support use a free online MT system to translate instructions. Twenty-nine participants were recruited and, at this phase of the research, divided into 'source' and 'machine translated' groups. The criteria for selecting the participants included being a native speaker of the target language, not having yet used the online storage services, while also being computer literate. The participants were asked to read the instructions and perform the tasks while their eye movements were being recorded. After they had finished the tasks, they answered a satisfaction questionnaire. The result of this first phase compared the machine-translated group against the source group, and found significant difference for goal completion, efficiency and user satisfaction between the source and the MT output. In the second phase of the study, Doherty and O'Brien (2014) analyse the results according to target languages compared to the source. The results show that the raw MT output scores lower for usability measurements, requiring more cognitive effort for all target languages when compared with the source language content. The target language Japanese (unsurprisingly, given its known difficulty as a target language for MT) scored lowest in terms of usability when compared to the other

target languages. In terms of satisfaction measured via the post-task questionnaire, English source content attained the highest ratings and, regarding the target languages, the raw MT output was still deemed usable, especially for Spanish.

Another study worth mentioning is Stymne et al. (2012) who present a preliminary study using eye tracking as a complement to MT error analysis. In this methodology, although the main focus is to identify and classify MT errors, a comprehension task is also applied. Four short texts from the Europarl corpus were selected and translated from English into Swedish via three different Moses-based MT systems trained on the same corpus: one trained with a larger number of sentences (*Large*), one trained with smaller number of sentences (*Small*), and the last one (*Comp*) with the same amount as the Large system, but with a compound processing module. The human reference translations were also used. Twenty-two native speakers of Swedish with good command of English, were instructed to read the texts for comprehension, with no time constraints, and asked to answer three multiple-choice questions about the text content, where participants also judge their confidence ratings for those multiple-choice questions. They were also provided with three perception questions, whereby they were asked to judge the fluency of the text, their own perceived comprehension of the text, and the perceived amount of errors in the text on an 8-point scale. The results for the comprehension text show that the number of correct answers for the reading comprehension is higher for the 'Large' system than for the human reference, but confidence scores are lower. For the perception questions, the human translation scored better than all the MT options. For both perceived and actual reading comprehension questions, the Large system is best, followed by Comp and, finally, Small. Regarding gaze data, MT errors have both longer gaze times and more fixations than correct passages, and average gaze time is dependent on the type of errors – which may suggest that some error types are more disturbing for readers than others. This result is corroborated in Aziz, Koponen and Specia (2014) who found that there is a connection between longer post-editing times when passages involve verbs, and higher number of edits when passages involve nouns.

Klerk et al. (2015) present an experimental eye-tracking usability test with human text simplification and machine translation (for both the original and

simplified versions) of logic puzzles. Twenty native speakers of Danish were presented with 80 different logic puzzles and asked to solve and judge the puzzles while having their eye movements recorded. The results demonstrated a greater number of fixations on the MT version of the original text (with no simplification). Regarding task efficiency, results show that participants were less efficient when using the MT version of the original puzzles; however, the simplified MT version seemed to ease task performance when compared to the original English version. The study, however, does not consider any type of post-editing.

The main drawback of Stymne's study is similar to the studies discussed in Section 2.6.4: it does not take into consideration all the aspects of usability, being limited to number of correct answers and gaze time. The studies of Doherty and O'Brien (2012, 2014), however, try to take into consideration all the aspects of usability as defined by ISO to include efficiency, effectiveness and satisfaction; therefore, the methodology seems suitable to investigate users' tolerance to different types of translation. This research intends to apply similar methodology in order to measure the usability for source, raw MT as well as post-edited texts.

2.8 Conclusion

The main goal of this chapter was to provide an overview on how TQA has been viewed and applied within translation studies and the translation industry. The Chapter discussed the gap existent between academia (where the focus is mainly on the theory and pedagogy of translation quality) and the translation industry (where the focus is on developing criteria bounded by internationally agreed standards) and the fact that no single standard for TQA has been achieved yet. A description and analysis of alternative ways of measuring quality that has emerged with the introduction of machine translation systems, such as automatic metrics and post-editing effort has also been presented, as well as how the end user is mostly disregarded when assessing translation quality and so, the acceptability of translated and source content is not regularly formally measured. The discussion also mentions the attempts that have been made by TAUS, QTLaunchpad, and QT21

in order to move to a more dynamic quality model that can take into consideration different views of translation quality (DQF and MQM models) including the view of end users as suggested by the broad definition of translation by Koby et al. (2014) and the UCT approach. The Chapter provides a description of the UCT approach which claims that end users should be involved in all stages of the translation process because the end user's feedback can be implemented in the translation process; and presents the few works that have attempted to measure the usability of translated texts with and without the help of eye tracking techniques and, therefore, the need for assessing the acceptability (usability, quality and satisfaction as per the definition used in this research) of MT and PE translations was identified. Measuring the acceptability of translated and source text allows for the identification of the impact that the translation might have on the end user.

Next, Chapter 3 introduces the Rationale that guides this research project, as well as the motivation and research questions.

Chapter 3 – Rationale

In the previous chapter, the issue of translation quality assessment (TQA) in academia and industry was discussed. The existing gap between the two areas was examined as well as how recent efforts have been made in order to bridge this gap by attempting to standardise TQA. The discussion also emphasised how the end user of translations has been largely disregarded when assessing translation quality and how little is known regarding how end users engage with raw machine-translated and post-edited texts.

As the main goal of this study is to investigate the acceptability for end users of raw and post-edited MT, it draws on the user-centered translation approach which emphasises the importance of considering the end user consistently throughout the translation process, including evaluation.

This chapter starts by defining the concept of acceptability and its three elements: usability, quality and satisfaction (Section 3.1), and examining the complex interrelationship between them within existing literature, highlighting consistencies and differences that need to be addressed, before providing combined definitions that are applicable to this study. The motivation for the present work is discussed in Section 3.2, regarding the choice of evaluating MT and PE in the industry in terms of acceptability (usability, quality and satisfaction) as well as the motivation for evaluating the source content in the same terms. The chapter closes with an explanation of the research questions and hypotheses (Section 3.3) that provide the rationale for the subsequent chapters and, indeed, the dissertation as a whole.

3.1 Operationalising Acceptability

The term ‘acceptability’ has been used in various fields such as linguistics, text linguistics, translation, and also in the field of human-computer interaction to refer to the level of acceptance of the end user (also reader, user, receiver, etc.) regarding language, a text, or a product.

In linguistics, Chomsky (1969), in his work on the theory of syntax, uses the term 'acceptable' to refer to "utterances that are perfectly natural and immediately comprehensible without paper-and-pencil analysis, and in no way bizarre or outlandish" (Chomsky 1969, p.10). The author separates the term 'acceptable' from 'grammatical' and states that acceptability is "a concept that belongs to the study of performance" - where performance relates to the "actual use of languages in concrete situations" (ibid., p.4), whereas grammaticality "belongs to the study of competence" (ibid., p.11). In his view, acceptability is a matter of degree(s) and can be specified through various operational tests.

In text linguistics, acceptability has been used as one of the standards of textuality defined by De Beaugrande and Dressler (1981), in which acceptability concerns "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" (De Beaugrande and Dressler 1981, p.7, emphasis in original), that is, the text should establish useful or relevant information such that it is 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" (ibid., p.113). Although Chomsky's definition of acceptability is concerned with the speaker-hearer, some similarity can be drawn between his and De Beaugrande and Dressler's concept, since one can argue that a sentence that is "bizarre or outlandish" will require some level of "tolerance towards [its] disturbances" and, therefore, a sentence will be more acceptable when it contains fewer disturbances.

Acceptability has also been used in the translation field. In Toury (1995), acceptability and adequacy are part of a paired concept of translation norms, in which acceptability relates to the adherence to target culture norms whereas adequacy relates to equivalence to the source text. Puurtinen (1995, p.230) states that acceptability in translated children's literature can be determined by the readability and speakability levels of a text, conformity to linguistic norms, and conformity to the expectations of the readers. For the author, there are different types of acceptability and, therefore, "a more complex, flexible concept, which allows of such heterogeneity" is needed (ibid.). These two definitions are similar in

claiming that, to be acceptable, a translation should concern the target culture norms and, therefore, the reader's expectation.

In Van Slype (1979, p.92), the concept of acceptability is used to assess the quality of machine translation and is defined as "a subjective assessment of the extent to which a translation is acceptable to its final user" that "can be effectively measured only by a survey of final users" (ibid., p.13), although he does not specify the type of survey questions to be put to the user. For Van Slype, measuring the acceptability of MT has several advantages, including: the use of simple criterion (i.e. the text is either acceptable or not); the judgement is made by the end user (i.e. the one the translation is done for); the measurement of acceptability relates to the actual purpose of the procedure (i.e. acceptance or not of the translated text by the user), and not to an intermediate or partial aspect (intelligibility, fidelity, etc.). This view of acceptability contrasts with previous definitions of the term – which claimed that acceptability can be seen as a "matter of degree" (Chomsky 1969) and a "complex concept" (Puurtinen 1995) – because this view holds that only a final survey with end users could assess the acceptability of translation. Nonetheless, several studies attempted to measure the acceptability of MT, drawing on Van Slypes' definition.

Coughlin (2003) uses the term acceptability to define integrated evaluation criteria, which consist of measuring comprehensibility, grammar and accuracy. In this study, participants were given a 1 to 4[14] scale to judge the acceptability of MT and human translated texts. It is interesting to note that, even though the author uses the term 'acceptable' in the questionnaire, the definitions for each heading include terms such as 'style', 'accurate', 'perfect', 'comprehensible'. The author claims that this approach exempts her from having to determine the importance of

[14] The scale consists of : "1= Unacceptable: Absolutely not comprehensible and/or little or no information transferred accurately; 2= Possibly Acceptable: Possibly comprehensible (given enough context and/or time to work it out); some information transferred accurately; 3= Acceptable: Not perfect (stylistically or grammatically odd), but definitely comprehensible, AND with accurate transfer of all important information; 4= Ideal: Not necessarily a perfect translation, but grammatically correct, and with all information accurately transferred" (Coughlin 2003, p.84).

fluency over adequacy (and vice-versa), since “raters balanced these different characteristics as they saw fit” (ibid., p.64).

Lassen (2003, p.XV) defines acceptability to include “grammatical acceptability as well as stylistic acceptability” when investigating the attitudes of users to the accessibility and acceptability in technical documentation via an offline survey. The author found that the term ‘acceptable’ is often problematic and is understood differently by some respondents and states that acceptability “is an ambiguous notion that may imply grammaticality to some respondents, while it may imply stylistic acceptability to others” (ibid., p.81).

Roturier (2006, p.4) bases his definition of acceptability on De 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” (ibid., p.157). The author acknowledges Van Slype’s statement which acceptability can only be measured via a survey with end users, and concludes that it is “essential that the evaluation of documents is performed by genuine users of such documents to maximise the ecological validity of the study” (ibid., p.149).

Van Slype’s concept of acceptability, supported by Coughlin (2003), Lassen (2003) and Roturier (2006), where only a survey with final users can measure acceptability, does not comply with the view of acceptability in this research. Acceptability is therefore conceived, as per Puurtinen’s and Chomsky’s definitions, as a complex concept, consisting of various degrees and that can be measured using a variety of different 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 operationalise acceptability.

Acceptability is also associated with human-computer interaction (HCI). For Nielsen (1993, p.24), 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”. In his model of

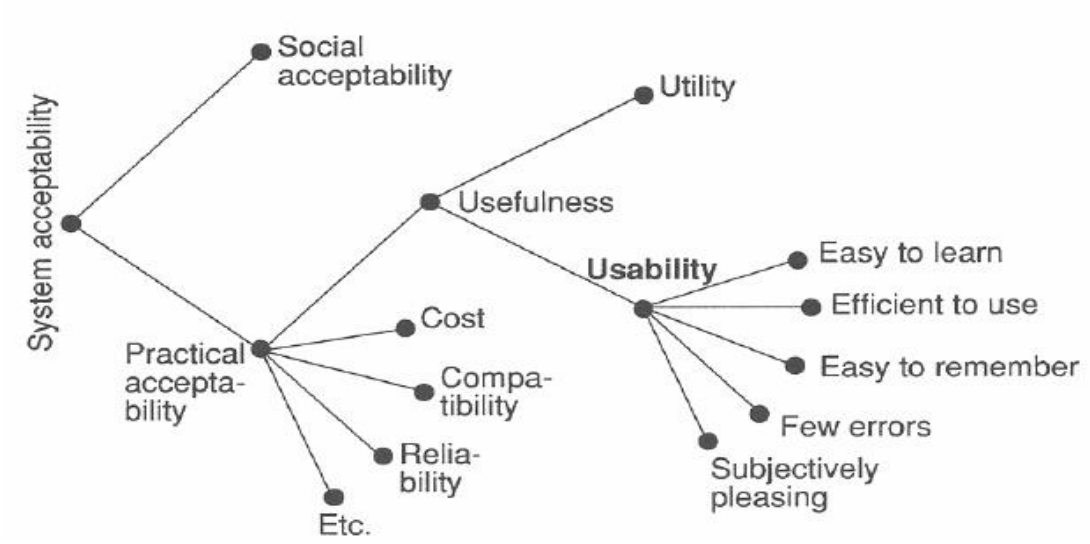


Figure 3:1 - Nielsen's System Acceptability Model (Nielsen 1993)

system acceptability (see Figure 3:1**Error! Not a valid bookmark self-reference.**), the author divides the "overall acceptability" of a computer system into "social acceptability" (highly socially desirable) and "practical acceptability". Practical acceptability consists of concepts such as cost, reliability, etc., as well as usefulness, which is defined as "the issue of whether the system can be used to achieve some desired goal" (ibid.) and is further divided into 'utility' and 'usability'. Nielsen, therefore, considers usability to be a narrow concern of the system acceptability model. For Suojanen, Koskinen and Tuominen (2015), Nielsen's notion of social acceptability is close to that of Toury's (1995) concept system in which acceptability relates to adherence to target culture norms and, moreover, acceptability, utility and usability from Nielsen's model are close to the ideas expressed in skopos theory (ibid., p.16) (see Chapter 2, Section 2.2 for a description of skopos theory). However, the authors claim that if an attempt was made to draw a model from the translation point of view, Nielsen's acceptability model would look slightly different. This is due to the fact that Nielsen's model classifies social acceptability far from the centre of usability (see Figure 3:1), and, in contrast, as translation studies has focused on the effect of target culture norms, acceptability is considered "particularly relevant because of the intercultural elements inherent in translation" (ibid.)

Figure 3:2 - Acceptability Model

This research focuses on acceptability as per Nielsen's acceptability model, where acceptability is composed of various categories. It complies with De Beaugrande and Dressler's concept of acceptability, in which acceptability refers to the relevance of a text for 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). And finally, the study brings the acceptability concept closer to the concept of usability and aims to measure the acceptability of machine translated instructional content via usability, satisfaction and quality (Figure 3:2). Applying Nielsen's model to translation, a user will find a translation (MT, PE or HT 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.

3.1.1 Usability

The term usability is usually associated with HCI. It was introduced in the 1980s to replace the term 'user friendly', which was considered overly vague (Bevan, Kirakowski and Maissel 1991). For Bevan, Kirakowski and Maissel (1991), usability relates to the interaction of the user with the product (or system) and only by assessing user performance, satisfaction and acceptability, is it possible to measure

usability accurately. Moreover, the authors state that a product is not *necessarily* usable or unusable, "but has attributes which will determine the usability for a particular user, task and environment" (ibid., p.654).

Usability research expanded from the engineering field - where it traditionally focused on studies of user interfaces - into other user-centred research, which include more abstract areas (Suojanen, Koskinen and Tuominen 2015). For Suojanen, Koskinen and Tuominen (ibid., p.14), usability is "ultimately about the *user's* relative experience of the success of use" [emphasis in original].Therefore, "almost any human activity can be studied from the point of view of usability" (ibid.) in which "we can all be perceived as users" (ibid., p.33).

According to Johnson, Salvo and Zoetewey (2007), usability research in technical communication had a turning point during World War II[15], with more research being conducted in order to understand how people read texts and apply what is learned to use technology (Suojanen, Koskinen and Tuominen 2015). Subsequently, usability gained a definition in an ISO standard, first in the 9241 *Ergonomics of human-system interaction* standard, and later in the ISO/TR 16982:2002 *Ergonomics of human-system interaction - Usability methods supporting human-centred design* standard - which provides information on human-centred usability methods that can be used for design and evaluation. In the 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, Koskinen and Tuominen (ibid.), however, point out that the concept of usability is not broadly used in the translation field and the few studies that research usability are predominantly limited to technical translation. Byrne (2006, 2012) is one of the few examples of research on usability in the translation field. Byrne's 2006 book, *Technical Translation. Usability Strategies for Translating Technical Documentation,* builds upon his doctoral thesis (Byrne 2004) which investigates whether the use of Iconic Linkage[16] improves usability in user guides.

[15] During WW II, soldiers were being taught to use heavy machinery with textbooks in a classroom. The results of those practices were often disastrous leading to injuries or death.

[16] "The use of the identical wording to present the same information recurring in a text" (Byrne 2004, p.1).

Although his 2004 study experiments only in the source language, the author highlights that, from the point of view of translation training, translators would benefit from the understanding of usability and technical writing since technical translators are "a specific type of technical communicator" (ibid., p.262). In his book (2006), the author focuses on usability strategies for translating technical texts and analyses in what ways translators can improve usability during the translation process. Byrne, again, asserts the importance of Iconic Linkage usage – which can be defined as the repetition of target language with semantically identical but non-isomorphic forms – in the translation process. Usability is, then, defined by Byrne (2014) 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" (ibid., p.201).

Suojanen, Koskinen and Tuominen (2015) claim that because usability research originates in the technology world, in which there is a product, and this product has a user, the term product most likely refers to a concrete technical device. However, texts can also be seen as a product and therefore, readers as users. Furthermore, the authors affirm that, even though it is easy to link usability and user to technical texts - as typically their purpose is to make users act according to clearly defined aims - usability research does not pertain to technical texts only, although the idea of usability is more easily "applied in some genres than others" (ibid., p.3).

This research focuses on usability as outlined by Suojanen, Koskinen and Tuominen (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" (ibid., p.14), together with Byrne's view of the extent to which users find this experience difficult or easy. For carrying out the experiments, the ISO definition for usability is adopted and, therefore, the concepts of effectiveness, efficiency and satisfaction are used. Chapter 4 describes in detail how those concepts are measured.

3.1.2 Quality

As discussed in Chapter 2, how best to assess the quality of translated texts is a controversial topic in the translation field. While TQA in the industry is mostly based on error typology models, academia mainly focuses on the theory of translation quality. Authors disagree on what the definition of translation quality should be and, moreover, in the translation industry, the tendency is to apply each company's own working definition of quality.

The present research uses the term quality in relation to two experiments: one performed with the translated content, and the other performed with the source content. For the translated content, quality assessment involves a TQA questionnaire answered by professional translators. As the goal of this study is to measure the acceptability of *translated enterprise content*, it stands to reason that, in order to ensure ecological validity of the quality experiment, the translated texts used in this research need to be also assessed by the regular method our industry partner applies to their content, and therefore, quality is measured via a TQA questionnaire (the details of the questionnaire are discussed in Chapter 4, Section 4.3.3.3). In this approach, translated texts need to present sufficiently few errors in order to be considered 'good enough' to be published on the company's website. As this research is also interested in correlations between target text acceptability and source text quality, the quality of the source content is also assessed for comprehension, readability, and complexity. The quality assessments are comprehensively described in Chapter 4.

3.1.3 Satisfaction

As outlined in Section 3.1.1, 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.

ISO has defined satisfaction as the "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 ISO's as well as Rubin and Chisnell's definition of satisfaction presented above. Moreover, Byrne's view on satisfaction also complies with the objectives of this research, whose goal is to measure user's subjective reactions, opinions, perceptions and attitudes towards the translated texts. Satisfaction is measured via three approaches: i) user opinions right after performing a task, ii) through professional translators, when evaluating the quality of the translation via a TQA questionnaire (see Section 4.1.2) and, iii) end users of the company's product via a live web survey displayed on the company's website. The methodology for measuring satisfaction is described in detail in Chapter 4.

3.2 Research Questions and Hypothesis

This research is driven by the following over-arching research question:

> **RQ**: What factors influence acceptability levels of a machine translated text for the end user?

Three factors form part of our hypothesis: **Post-editing Level**, **Language** and **Source Content** quality influence acceptability levels. Acceptability is composed of different constructs, defined here by usability, quality and satisfaction, which in turn, guide the experiments performed. Therefore, the research questions are separated here first by the three factors mentioned above and by the constructs of acceptability.

Factor One: Post-editing Level

In this research, two post-editing levels are considered:

A) No post-editing or PEZero (hereafter 'PEz'): when the content is translated solely by a machine translation system, that is, raw machine translated output. This output may contain issues such as grammatical, syntactic and terminology errors, that are expected to affect the acceptability of the output by end users.

B) Professional 'light' post-editing (hereafter 'PEp'): when the content is translated by a machine translation system and its output is modified by a human translator, that is, a post-edited version of this output. This post-edited output is naturally thought to have fewer problems when compared with raw MT output, but may still affect acceptability levels among some end users. The guidelines provided for the translators for the light post-editing is described in Chapter 4, Section 4.3.3.1.

In order to test whether the factor Post-editing Level influences the acceptability level, three experiments are implemented to focus on usability, quality and satisfaction. Two sub-questions are posed for the Post-editing Level and usability:

RQ1: Does Post-editing Level have an effect on usability?

The null hypothesis for Post-Editing Level and usability can be expressed as follows:

H1.1$_0$: There is no difference in levels of usability between raw MT output (PEz) and lightly post-edited MT output (PEp).

The alternative hypotheses can be expressed as follows:

H1.1$_1$: Higher levels of usability are evident for the PEp texts.

H1.1$_2$: Despite lower levels of usability, the PEz texts still allow for goal completion.

Two sub-questions are posed for Post-editing Level and satisfaction:

RQ2: Does Post-editing Level have an effect on satisfaction?

The null hypothesis for Post-editing Level and satisfaction can be expressed as follows:

H1.2$_0$: There is no difference in levels of satisfaction between raw MT output (PEz) and lightly post-edited MT output (PEp).

The alternative hypotheses can be expressed as follows:

H1.2$_1$: Higher levels of satisfaction are evident for the PEp texts.

For Post-editing Level and quality, one sub-question is posed:

RQ3: Does the quality evaluation of Post-editing levels PEz and PEp, performed by professional evaluators reflect the results from the empirical usability and satisfaction experiments?

The null hypothesis for Post-editing Level and quality can be expressed as follows:

H1.3$_0$: The quality evaluation does not reflect the results from the empirical usability and satisfaction experiments.

Factor Two: Language

Languages also play an important role in our hypothesis. The fact that some languages are more challenging for MT may have an effect on the usability levels for post-editing level PEz. It is known that every change in word order is problematic for MT systems, especially if it pertains to topicalisation of non-subject arguments (such as German). In addition, languages for which word segmentation cannot be easily established (such as Japanese and Chinese) are problematic as words are the basis of most translation models. Birch, Osborne and Koehn (2008) found that reordering, morphological complexity of target language, and historical relatedness of the two languages are strong predictors of MT performance. Additionally, if anecdotally, the industry partner whose content and MT system were used for this

research identified German, Japanese and Chinese as some of the most challenging target languages among a broad spectrum of languages into which they (machine) translate.

Another aspect is that users of different languages may also have a different threshold of tolerance for translation disturbances. As discussed previously, industries have different thresholds/approaches for dealing with different countries regarding translation. For example, while Brazilian Portuguese speakers might accept reading marketing content that was just machine translated, French users might demand the same content with a higher level of quality.

Taking language as a factor that affects acceptability levels, research questions are considered for usability, satisfaction and quality.

One sub-question is posed for Language and usability:

> **RQ4:** How do different target languages compare in terms of usability for both PEp and PEz content?

The null hypothesis for Language and usability can be expressed as follows:

> **H2.1$_0$**: There are no differences in usability levels of PEp and PEz content for German, Simplified Chinese and Japanese.

One research question is posed for the factor Language and satisfaction:

> **RQ5:** How do different target languages compare in terms of satisfaction for both PEp and PEz content?

The null hypothesis for Language and satisfaction can be expressed as follows:

> **H2.2$_0$**: There are no differences in satisfaction levels of PEp and PEz content for German, Simplified Chinese and Japanese.

For Language and quality, one sub-question is posed:

RQ6: Does the quality evaluation of the translated languages performed by professional evaluators reflect the results from the empirical usability and satisfaction experiments for Language?

The null hypothesis for Language and quality can be expressed as follows:

H2.3$_0$: The quality evaluation does not reflect the results from the empirical usability and satisfaction experiments for Language.

We expect that there will be differences and that German might show higher levels of usability, quality and satisfaction for PEz and PEp content over Chinese and Japanese.

Factor Three: Source Content

As presented in Section 3.2.3, source content evaluation does not follow a standard practice among companies. Several companies have professional linguists that perform spot checks on a percentage of the text while other companies use automatic validation checks. Sending faulty source content back to the authoring team is not a regular practice and, at times, companies expect that the translator will spot the errors when translating and even - in some cases - correct the errors in the source content. If the content is machine translated, those errors may cause similar or other errors in the translated version.

Taking source content as a possible factor that affects machine translation acceptability levels, research questions are considered for usability, quality and satisfaction.

One sub-question is posed for the Source Content and usability:

RQ7: How does usability of Source Content compare with usability of the translated content (PEp and PEz)?

The null hypothesis for Source Content regarding usability can be expressed as follows:

> **H3.1$_0$**: There are no differences in usability levels of Source Content when compared to PEp and PEz content for German, Simplified Chinese and Japanese.

One sub-question is posed for the Source Content and satisfaction:

> **RQ8:** How does satisfaction with Source Content compare with satisfaction with translated content (PEp and PEz)?

The null hypothesis for Source Content regarding satisfaction can be expressed as follows:

> **H3.2$_0$**: There are no differences in satisfaction levels of Source Content when compared to PEp and PEz content for German, Simplified Chinese and Japanese.

For Source Content and quality, one sub-question is posed:

> **RQ09:** Does the quality evaluation of the Source Content reflect the results from the empirical usability and satisfaction experiments for Source Content?

The null hypothesis for Source Content and quality can be expressed as follows:

> **H3.3$_0$**: The quality evaluation does not reflect the results from the empirical usability and satisfaction experiments for Source Content.

The alternative hypotheses for the Source Content can be expressed as follows:

> **H3.1$_1$**: Higher levels of usability and satisfaction are visible for the Source when compared to the PEp and PEz texts.

H3.1$_2$: The same levels of usability and satisfaction are visible for Source, PEp and PEz texts.

In summary, this study is conducted under the main research question "*RQ: What factors influence acceptability levels of a machine translated document for the end user?*" for which we examine three factors: *Post-Editing Level, Language* and *Source Content*. The RQ is, therefore, broken into multiple sub-RQs as shown in Table 3:1.

In order to test the hypotheses, different but complementary methods for measuring usability, quality and satisfaction have been implemented and are discussed in detail in Chapter 4.

Research Questions	Null Hypothesis	Factor this RQ address
RQ1: Does Post-editing level have an effect on usability?	**H1.1$_0$**: There is no difference in levels of usability between raw MT output (PEz) and lightly post-edited MT output (PEp).	**Post-Editing Level**
RQ2: Does Post-editing Level have an effect on satisfaction?	**H1.2$_0$**: There is no difference in levels of satisfaction between raw MT output (PEz) and lightly post-edited MT output (PEp).	
RQ3: Does the quality evaluation of Post-editing levels PEz and PEp, performed by professional evaluators reflect the results from the empirical usability and satisfaction experiments?	**H1.3$_0$**: The quality evaluation does not reflect the results from the empirical usability and satisfaction experiments.	
RQ4: How do different target languages compare in terms of usability for both PEp and PEz content?	**H2.1$_0$**: There are no differences in usability levels of PEp and PEz content for German, Simplified Chinese and Japanese.	**Language**
RQ5: How do different target languages compare in terms of satisfaction for both PEp and PEz content?	**H2.2$_0$**: There are no differences in satisfaction levels of PEp and PEz content for German, Simplified Chinese and Japanese.	
RQ6: Does the quality evaluation of the translated languages performed by professional evaluators reflect the results from the empirical usability and satisfaction experiments for Language?	**H2.3$_0$**: The quality evaluation does not reflect the results from the empirical usability and satisfaction experiments for Language.	
RQ7: How does usability of Source Content compared with usability of the translated content (PEp and PEz)?	**H3.1$_0$**: There are no differences in usability levels of Source Content when compared to PEp and PEz content for German, Simplified Chinese and Japanese.	**Source Content**
RQ8: How does satisfaction with Source Content compare with satisfaction with translated content (PEp and PEz)?	**H3.2$_0$**: There are no differences in satisfaction levels of Source Content when compared to PEZ and PEP content for German, Simplified Chinese and Japanese.	
RQ09: Does the quality evaluation of the Source Content reflect the results from the empirical usability and satisfaction experiments for Source Content?	**H3.3$_0$**: The quality evaluation does not reflect the results from the empirical usability and satisfaction experiments for Source Content.	

Table 3:1 - Research Questions, Null Hypothesis and Factors

Chapter 4 – Methodology

This chapter addresses the methodology followed for the assessment of acceptability. As discussed in Chapter 3, the acceptability model defined in this research consists of three elements: usability, quality and satisfaction and, therefore, different and yet complementary experiments were performed in order to assess those elements.

As can be seen in Figure 4:1, the assessment of quality and satisfaction differ for the source content and the translated content. For the source content, quality is measured with the help of two tools (see Section 4.2.2.2); whereas for the translated content, quality is evaluated by professional moderators via a TQA questionnaire (i.e. adequacy, fluency, terminology). For both source and translated content, satisfaction is measured via a web survey with user ratings and also end-user ratings performed after the usability experiments via a post-task questionnaire. However, the translated content is also assessed for satisfaction by professional moderators with the same TQA used in the quality assessment. The remainder of this chapter is structured as follows: First, Section 4.1 describes a pilot experiment that was performed in order to test the design for the experiments. Following that, the methodology applied for the experiments with the source content (4.2) and with the translated content (4.3), which includes the German, Simplified Chinese and Japanese languages, is presented.

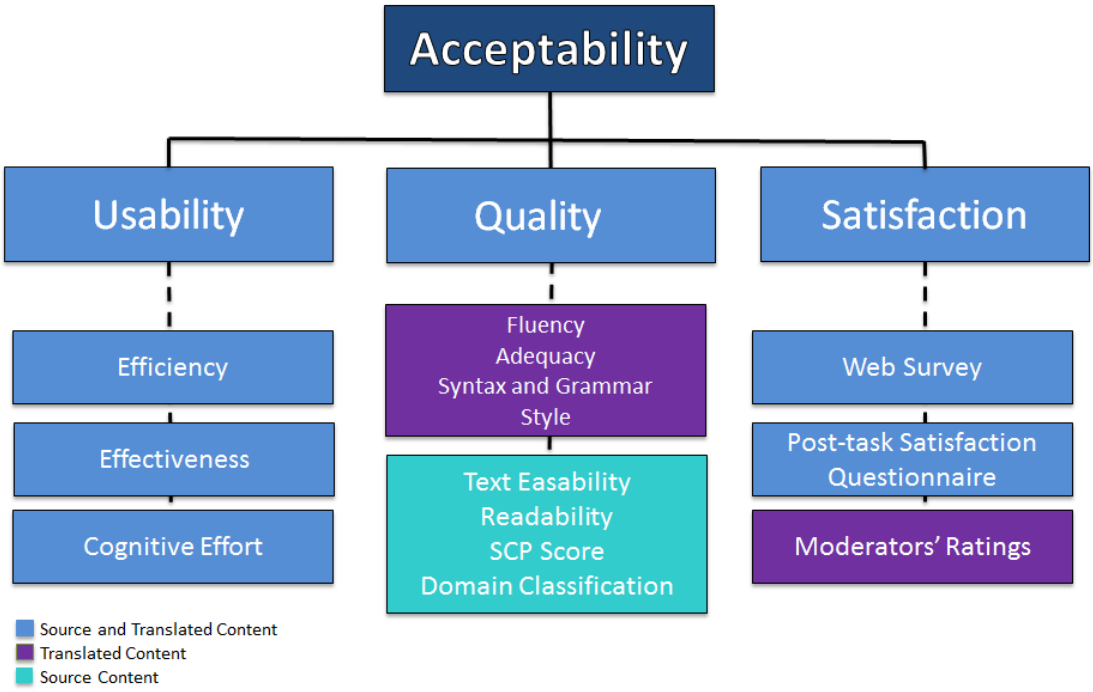


Figure 4:1 - Acceptability Model - Measures

4.1 Pilot Study

In order to identify issues and test the design, a pilot experiment was performed in collaboration with an industry partner who produces a software security product. The pilot took place at the Federal University of Minas Gerais, Belo Horizonte, Brazil in November 2013. The results of this pilot were published at the EAMT (European Association for Machine Translation) conference in 2014 (see Castilho et al. 2014).

Participants

Eighteen native speakers of Brazilian Portuguese were recruited from the student body of the Federal University of Minas Gerais, Belo Horizonte, Brazil. It was ensured that participants had no previous experience of this particular security product so that previous knowledge could not be used to compensate for poor quality of the machine translation output (Moravcsik and Kintsch 1995). The participants were randomly assigned to one of two groups: Group 1 used the raw machine translated output and were asked to follow the instructions while Group 2

read and followed the post-edited instructions. Neither group knew that the texts they were reading had been translated. Participants were seated at the eye tracker and were informed that they would be presented with some instructions on the left-hand side of the screen and a software product on the right hand side in which they had to perform five tasks as per the instructions. The tasks involved setting up an automatic cleaning schedule, setting parental controls, creating a vault, shredding files and deleting a vault. Participants were instructed not to reposition any of the windows relating to the software product or the instructions, so as to facilitate eye-tracking analysis.

Content

As mentioned, the security software product controlled for viruses and allowed for the setting of parental controls. Some instructional content in English on how to configure features of this product was selected. The total number of words in the source content amounted to 594. This content was machine translated into Brazilian Portuguese using Microsoft's Bing engine – as the company-specific MT engine could not be used at the time. Brazilian Portuguese was selected for this study as it was part of a Brazil/Ireland research collaboration project. The raw machine translated output was post-edited by a native speaker of Brazilian Portuguese who has an undergraduate degree in linguistics and literature and a Master's degree in natural language processing and human language technology, who also conducted research previously on post-editing. The guidelines adhered to during post-editing were those of TAUS for the level "fit-for-purpose" (TAUS: online), which meant that edits were carried out when terminology did not conform to the client-specific glossary and grammatical errors were fixed. No edits were implemented for purely stylistic reasons and the focus was on accuracy and comprehensibility. To measure how much post-editing was performed, an automatic evaluation comparing the post-edited version against the MT output was conducted. An average HTER score (see Section 2.6.1) of 0.20 was observed, which indicates that post-editing was of a light nature.

Procedure

The main methodological approach was to record time, reading and task completion data via the eye tracker. Both groups were given a warm-up task where they were asked to read a text in Brazilian Portuguese for comprehension; the text came from Wikipedia and explained the concept of virus checking. Fixation data gathered during this reading exercise were used as a baseline measurement for 'reading for comprehension' in Brazilian Portuguese among participants. There were no significant differences in reading data for the warm-up task between the two groups. Once they had completed their tasks they responded to a questionnaire, which addressed the construct satisfaction. Two participants (one from each group) appeared to be outliers in terms of several of the fixation measurements and so, their data was removed from each group.

Measuring Usability and satisfaction

The ISO/TR 16982 definition for usability was adopted, which, characterizes usability as effectiveness, efficiency, and satisfaction (ISO 2002). When this definition is divided into its component parts, it allows us to measure different aspects of usability using a variety of methods. Effectiveness is measured through goal completion, that is, how successful the users were at accomplishing tasks documented in the instructions measured by observing the user interactions as recorded by a Tobii T60XL eye tracker. Efficiency is measured as the number of successful tasks completed (out of all possible tasks) when total task time is taken into account. A second measure of efficiency is cognitive effort, i.e. how much cognitive effort is evident when users are reading the instructions and trying to complete their tasks? Cognitive effort is measured using typical indicators recorded via the eye-tracking apparatus, i.e. mean total fixation time, mean fixation duration, total fixation count, average visit duration and visit count. As discussed in Chapter 2, such fixation data are well established as indicators of cognitive effort (Rayner 1998, Radach et al. 2004). For example, the more fixations there are on a set of instructions, the more probable it is that the reader is having difficulties in processing the instructions.

Satisfaction is a measure of user satisfaction with the translated content and, by extension, the product itself. As satisfaction is a multi-faceted concept, it is measured by using a questionnaire with a Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). In the questionnaire, "satisfaction" is addressed using a number of statements: 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 be able to use the software again in the future without re-reading the instructions; 7. I would recommend the software to a friend or a colleague; 8. I would consider buying this product after participating in this experiment.

The results of the pilot study showed that light post-editing improves the usability of the texts translated from English into Brazilian Portuguese, thus providing a natural hypothesis that light post-editing may also improve the usability for languages that are more 'challenging' for MT (results are presented in detail in Castilho et al. 2014).

Lessons Learned from Pilot Study

The pilot showed that post-editing – even to the level of "fit-for-purpose" – adds value to machine translated content by increasing usability and satisfaction levels. Since the language used was Brazilian Portuguese, which is considered to be one of the more successful target languages for MT, the question whether this result could be replicated with languages that are considered to be "challenging" for MT arises.

As an exercise which directly feeds into the research undertaken in this thesis, the pilot also served as a means to generate and test the design of the experiments, the fundamentals of which have been suitably adjusted and applied to the languages of English, German, Japanese and Simplified Chinese. These adjustments apply to the categories of: design, metrics, participants, and questionnaire.

The pilot demonstrated that the design was appropriate because the participants were able to complete the task and motivated to do so. They understood the tasks which were theoretically relevant to them (virus checking,

setting up security on a PC etc.) The screen setup used in the pilot is replicated as it proved to be satisfactory for reading instructions and performing tasks without changing between windows. Therefore, for the main experiment, content that is relevant to the community of participants available for the eye-tracking experiment was needed, which, in the case of this research are university students.

In terms of participants, it was observed that some people are naturally not suitable for eye tracking and several recordings were not utilised as they have low percentages for recording quality.[17] Therefore, a greater number of participants is necessary since some of the recordings may be unutilised. In relation to metrics, it was noticed that the ones used in the pilot (fixation and visits) gathered a great amount of data that made it possible to arrive at the results. Therefore, the same metrics are used in the main experiment as well.

4.2 Source Content Experiments

The experiments performed with the English language addressed usability, quality and satisfaction – the elements which constitute our model of Acceptability. Analysing the source content (EN_Source) is an important part of the study since it allows us to identify potential problems in the source which could be automatically transferred to the translated text. In this section, the data collection and the analyses of the data for the source content experiments are described.

4.2.1 Participants

This section describes the participants of all the experiments performed for the source language. For clarity, this section is divided according to the experiments performed: usability, quality and satisfaction.

[17] The percentage is a rough estimate of the quality of the eye tracking in a recording.

4.2.1.1 Usability Experiments

Eight native speakers of English were recruited from the student and staff body of Dublin City University, Dublin, Ireland. Ethics approval was granted by the relevant university research ethics committee.[18] The participants were between 27 and 39 years old, five male and three female. Seven of them hold a post-graduate degree and one a bachelor degree.

Differently from the pilot, it was not ensured that participants had no previous experience of this particular product as it is a well-known product. Instead, we wanted to ensure they were literate in the software, that is, that they would be able to deal with the basic functions.

In order to measure cognitive effort by analysing the fixations and visits, it was decided that the quality of the eye tracker recording should have a percentage higher than 80% and, for this reason; one recording had to be excluded from this analysis. In total, seven recordings were used to measure cognitive effort. Note that for the efficiency and effectiveness, all eight participants' data was used since the data for time and goal completion (and satisfaction) could be measured (i.e. these measures did not rely on the quality of eye tracking data).

4.2.1.2 Satisfaction Experiments

4.2.1.2.1 Web Survey

As the web survey is displayed on the industry's partner website, the participants of the survey are real end users who look for help when using the company's office suite products. By collecting these ratings, it is possible to gain an indication of real end-user satisfaction.

4.2.1.2.2 Post-task satisfaction Questionnaire

The post-task questionnaire was displayed after the usability experiment in order to assess the participants' level of satisfaction after using the instructions to

[18] See Appendix A for Plain Language Statement and Informed Consent Form.

perform specific tasks. Consequently, the participants who answered the satisfaction questionnaire were the same eight native speakers of English who participated in the usability experiments (4.2.1.1).

4.2.2 Materials

4.2.2.1 Content

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

The articles describe features of the spreadsheet application as well as instructions on how to use such features and are published on the company's website.[19] The choice of the content is motivated by several factors: i) the easy access users have to this content online which would allow for a wide-scale survey on satisfaction; ii) the willingness of the company to provide the content; ii) the theme of the content being, somehow, instructional which allows for creation of tasks that users can perform during the eye-tracking experiments.

For the satisfaction experiments performed via the web survey, 140 articles were selected and published online.

For the usability, quality and satisfaction (post-task satisfaction questionnaire) experiments, six articles were chosen and eight tasks were created. In total, the corpus consisted of 540 words. Originally, the articles published online contain images of the software such as buttons, icons, etc., however, as the goal of the experiment is to measure the usability of the text; some of the artwork was removed from the text. It was made sure that only the art that was complementing the text and not what was needed for understanding 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:

[19] See Appendix B for the articles used.

1) Quickly change colors, 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 color
6) Insert an exploding pie chart
7) Insert a bar of pie chart
8) Hide comments and their indicators

Tasks 6 and 7 were created from the same article; therefore five articles were used to create six tasks. Tasks 4 and 8 were also created from the same article and were chosen because human translated versions were available in the target languages (DE, ZH and JP). As mentioned in Section 1.1.1, the HT versions were incorporated as two control tasks.

A short text about office suites was selected from Wikipedia and displayed for the participants before they started the tasks as a warm-up exercise. The text, which was displayed in English and contained 160 words, is also used for recording a reading baseline, that is, fixation count and duration would be recorded.[20]

4.2.2.2 Tools

This section describes the tools used for each experiment in this research project. The tools used for the English Language experiments for usability, quality and satisfaction are described. It is important to note that, as some of these tools will be used for all the languages evaluated in this research (EN, DE, ZH and JP), the tools are described in detail when the term first appears and then referred to in subsequent sections.

[20] See Appendix C for full text.

4.2.2.2.1 Spreadsheet Software

In collaboration with the industry partner, a spreadsheet application from the office suite to be used as the software for the usability experiment was selected. The application includes calculation and graphing tools and is extensively used to carry out data manipulations. The choice of this application is due to the fact that, as the office suite has more than 1.2 billion users, an application in which participants would be literate but not total experts needed to be chosen. For that, it was also decided to use the newest version of the software, 2013, as it was assumed fewer people would have used that version.

4.2.2.2.2 Eye Tracker Device

The device used in this experiment is a Tobii T60XL, a wide-screen eye tracker - 24 inch monitor- with a 60Hz sampling rate. It has high screen resolution, allowing for studies of detailed stimuli[21], which is essential to this experiment since the participants need to have a clear view of all the spread sheet 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 a mixed content stimuli was chosen (see Figure 4:2 for screen layout).

4.2.2.2.3 Source Content Profiler

Source Content Profiler (SCP) is a tool developed by the CNGL/ADAPT research group at Dublin City University. The tool allows for the classification of documents into various profiles by making use of a language model trained on the National British Corpus (NBC), and a domain classifier. When a text is uploaded, the tool displays an overall score (SCP score) which measures the quality of an input document - on a scale from 0 to 100, where the higher the score the higher the quality of the document – and allows for the identification of the amount of issues in the content. It then breaks down those issues into shallow features, such as:

[21] See http://www.tobii.com [Last accessed 07 May 2016]

- Word and sentence length and number
- Syntactic structure including grammar issues, number of sentences with unusual POS sequences and passive voice issues
- Spelling issues
- Terminology used
- Domain detection

The objective of using the SCP was to better understand the features of the selected content, as well as its level of difficulty.

4.2.2.2.4 Coh-Metrix

Coh-Metrix is a computational tool that measures cohesion and coherence for written and spoken texts (Graesser et al. 2004). Coh-Metrix analyses texts on over 200 measures of language and readability, and over 50 types of cohesion relations by using lexicons, part-of-speech classifiers, syntactic parsers, templates, corpora, latent semantic analysis, and other components. Coh-Metrix has been used to identify differences between spoken discourse and written text, differences between writing styles (McCarthy et al. 2006), as well as to predict the difficulty of reading texts for second language learners (Crossley, Greenfiel and McNamara 2008). Therefore, Coh-Metrix has been proven to be a powerful text analysis tool that is capable of assessing different content types. The main objective of using this tool is to identify the level of comprehension difficulty of the corpus used for the experiments and, consequently, identify whether problems with the source content (if any) may influence the acceptability of the translated content.

4.2.2.2.5 Web Survey

A web survey displayed on the industry partner's website for 140 articles (EN, DE, ZH and JP) gathered information on 'how useful' the content is for the end user. The online survey consisted of only one multiple choice question: *"Was this information helpful?"* (YES/NO). Unfortunately, the survey question could not be changed nor could a second question be added as it is standard for the company's

website. One important point to be mentioned here is the implications of collaborating with companies for academic research. While the lack of control over certain parts of research may be a drawback - such as the availability of content types or the phrasing of web survey questions and other legal matters, the benefits that come with the collaboration – such as great amount of content when a type is agreed on, professional translation and post-editing and the end user ratings for known software making the research closer to the real world problem - outweigh those drawbacks. While, of course, a more detailed survey would be desirable, evaluating the 140 articles by this metric provides an *initial* indication of satisfaction levels that are complemented with the eye-tracking experiments.

4.2.2.2.6 Post-task Satisfaction Questionnaire

A post-task questionnaire in English was presented to participants after the performance of the usability experiments, and consists of nine questions with a Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). For all 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.

English native speakers did not see question 8 since they were already using the original version of the instructions.

4.2.3 Procedure

In order to make the description clearer, we have divided this section into usability experiments (4.2.3.1), quality experiments (4.2.3.2) and satisfaction experiments (4.2.3.3).

4.2.3.1 Usability Experiments

4.2.3.1.1 Recruitment Survey

An online pre-participation survey was employed in order to collect participants' demographic information as well as their availability for the experiment. Participants were asked to answer questions such as:

- Gender
- Age
- Education level
- English proficiency
- Software usage, version and frequency of use

After answering the recruitment survey, participants were asked to select a date for the eye-tracking experiment which was held in DCU from April to December 2015.

4.2.3.1.2 Tasks

Upon arriving to do the experiment, participants were asked to read the Plain Language Statement and sign the Informed Consent Form. The researcher would explain what eye tracking is and how the experiment works. Participants were seated at the eye tracker and were instructed not to reposition any of the windows relating to the software product or the instructions, so as to facilitate eye-tracking

analysis. They were informed that they would be presented with a short text in their language and were instructed to read the text for comprehension. Fixation data gathered during this reading exercise was used as a baseline measurement for "reading for comprehension". After reading the text, the participants answered the following question:

Q. How often do you use this spreadsheet application?

- Every day
- Two/three times a week
- Once a week
- Once a month
- Never

It was decided to ask this question during the experiment and not in the pre-task survey so that the participants would not know beforehand that a spreadsheet application would be used as the software for the usability experiment. Upon beginning the experiment, the participants would see the instructions on the left-hand side of the screen and the software product on the right hand side in which they had to perform eight tasks as per the instructions. Each task was presented individually, that is, the participants would see the instructions for one task at a time. Figure 4:2 shows the screen layout.

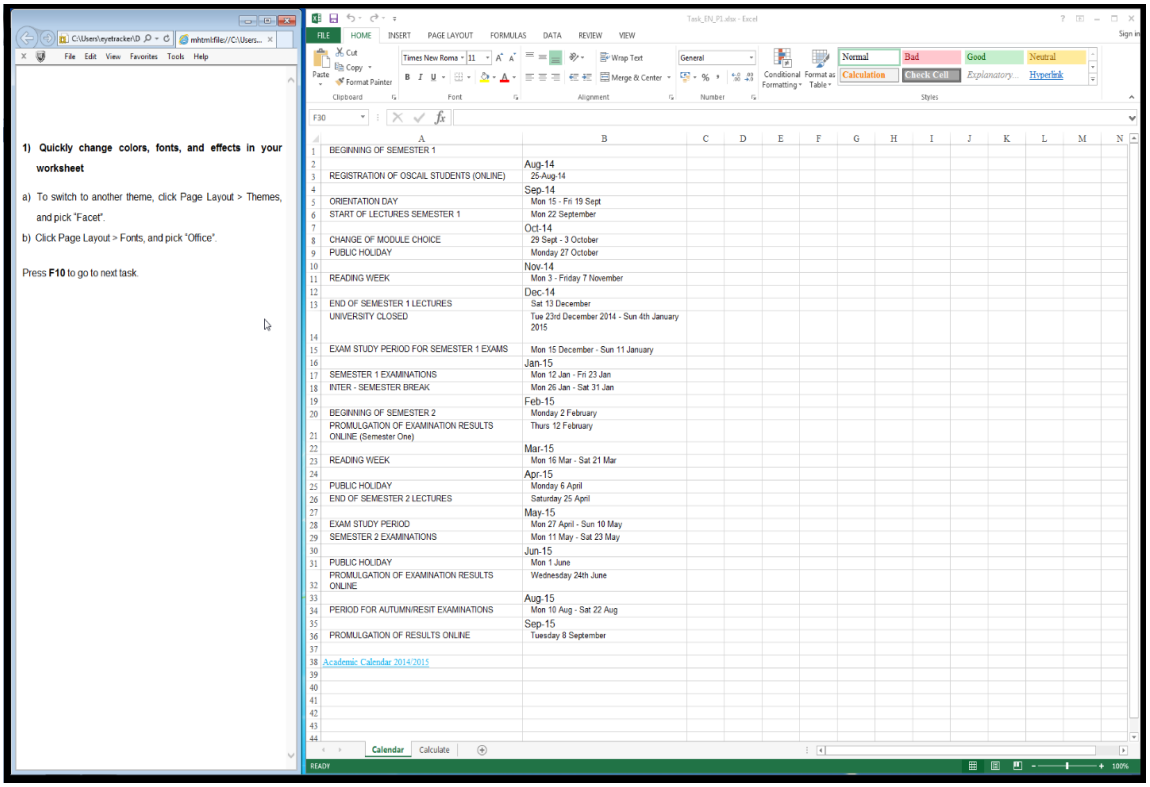


Figure 4:2 - Task Design

4.2.3.2 Quality Experiments

Two tools were used for the quality assessment of the English language: Source Content Profiler and Coh-Metrix. As described in Section 4.2.2.2.4, the aim of this experiment is to understand the features of the selected content, as well as its level of comprehension difficult in order to verify whether problems with the source content (if identified) may influence the acceptability of the translated content.

The SCP was used in May 2016 on the Adapt Centre website.[22] The English content (see 4.2.2.1) was uploaded to the user interface of the tool in a txt format and the results were displayed in percentages and charts.

The Coh-Metrix tool was also used in May 2016 on the tool webpage with the same content used for the SCP tool, with the exception that instead uploading a file to the tool, the text has to be copied and pasted in the tool web interface. The tool displays the results in a table on the website.

[22] The tool is available for members of the Adapt Centre.

4.2.3.3 Satisfaction Experiments

As satisfaction is a multi-faceted concept, it is measured via two different approaches for the English language experiments:

1. End-users' rating for satisfaction via a web survey (described in 4.2.2.2.5)
2. Post-task satisfaction questionnaire, performed after the usability experiment (described in 4.2.2.2.6)

4.2.3.3.1 Web Survey

As mentioned before, a web survey was displayed on the industry partner's website and gathered information on 'how useful' the content is for the end user. The English articles are published on the company's website and the period used to gather the ratings was from July to September 2015. This time period was chosen as it was the same time some of the machine translated articles were also online (see Section 4.3.3.4.1). The survey gathered 5 thousand ratings for the English language, which was the highest number of ratings for the web survey across languages.

4.2.3.3.2 Post-task satisfaction Questionnaire

As mentioned previously, a post-task questionnaire was designed to capture the users' satisfaction level after performing specific tasks with the EN_Source (also PEz and PEp) instructions. The questionnaire was displayed on the eye tracker screen after the completion of the last task of the usability experiments. Participants were asked to take all the tasks into consideration when answering the questionnaire. As seen in Section 4.2.2.2.6, the questionnaire consisted of 9 questions with a Likert scale from 1 (Strongly Disagree) to 5 (Strongly Agree) as multiple choice answers. The participants were required to click on one of the options.

4.3 Translated Content Experiments

The translated content assessed in this research consists of three languages: German, Simplified Chinese and Japanese. As for the source content, the

experiments performed with the translated content languages consisted of usability, quality and satisfaction. Two post-editing levels for each language were measured: one is the raw machine translation version, that is, no post-editing performed (PEzero - PEz); and the second is a lightly post-edited version of the raw MT output (PEprofessional - PEp). In this section, we describe the data collection and the analyses of the data for the translated content, for each language and post-editing levels.

4.3.1 Participants

4.3.1.1 Usability Experiments

The selection for participants for the usability experiments followed the same criteria used for the source content experiments (see 4.2.1.1), where participants were recruited from the student and staff body of Dublin City University, Dublin, Ireland.

4.3.1.1.1 German

Fourteen native speakers of German volunteered for the study. The participants were between 21 to 51 years old, six male and eight female. Nine participants hold a post-graduate degree and five have undergraduate degree.

As mentioned before, a threshold of 80% was set for recording quality in order to be able to measure cognitive effort by analysing the fixations and visits, and for this reason, one recording had to be excluded from this analysis. In total, thirteen recordings were used to measure cognitive effort (seven for the PEz and six for the PEp groups). Note that for the efficiency and effectiveness measures, the data for all fourteen participants was used since the data for time and goal completion (and satisfaction) was not impacted by eye tracking data quality.

4.3.1.1.2 Simplified Chinese

Twenty-one native speakers of Simplified Chinese were recruited for the study. Their age range was from 20 to 39 years and nine are male and twelve

female. Sixteen participants hold a post-graduate degree and five have an undergraduate degree.

Because of the quality of the recordings, six had to be excluded from the cognitive effort analysis. The number of excluded recordings may be due to the fact that many of the participants used glasses and contact lenses. In total, fifteen recordings were used to measure cognitive effort (seven for the PEz and eight for the PEp groups). Note that for the efficiency and effectiveness measures, the data for all twenty-one participants was used since the data for time and goal completion (and satisfaction) was not impacted by eye tracking data quality.

4.3.1.1.3 Japanese

Twenty-eight native speakers of Japanese volunteered for the study. The participants were between 18 to 56 years old, twenty were female and eight male. Eight participants hold a post-graduate degree; eleven hold an undergraduate degree and nine participants were exchange students, who came to the university to improve their English.

Regarding the quality of the recordings, the Japanese participants were the ones who presented the lowest percentage. For that reason, fourteen recordings had to be excluded from the cognitive effort. We speculate that, apart from participants who used glasses and lenses, the shape of Asian eyes may have had an impact on the quality. In total, fourteen recordings were used to measure cognitive effort (seven for each group). Note that for the efficiency and effectiveness measures, the data for all twenty-eight participants was used since the data for time and goal completion (and satisfaction) was not impacted by eye tracking data quality.

4.3.1.2 Quality Experiments

The quality experiments were performed by eighteen moderators from the company's language service provider - LSP (six for each language – DE, ZH and JP). According to the supplier, a moderator is slightly different from a translator because moderators have 'solid experience with reviewing and quality evaluation'.

4.3.1.3 Satisfaction Experiments

4.3.1.3.1 Web Survey

Similar to the source content experiment (see 4.2.1.2), the participants of the web survey are real end users who seek instructions on how to use the company's office suite products.

4.3.1.3.2 Post-task satisfaction Questionnaire

As for the source content (see 4.2.1.2), the post-task satisfaction questionnaire was displayed after the usability experiment for the translated content, and, therefore, the participants who answered the satisfaction questionnaire were the same native speakers of German, Simplified Chinese and Japanese who participated in the usability experiments.

4.3.1.3.3 Moderators' rating (TQA)

The ratings for satisfaction were given by eighteen moderators that performed the quality experiments (see 4.3.1.2).

4.3.2 Materials

4.3.2.1 Content

The selected corpus for the experiments was a machine translated version of the source content selected for the English language experiments (see Section 4.2.2.1). The texts were translated with the company's machine translation system and then lightly post-edited by professional translators who have experience working with the industry partner. The languages selected, as previously mentioned, were German, Simplified Chinese and Japanese, and were chosen for being known to be more problematic for machine translation systems. This aspect of those languages was also of concern for the industry partner as they consider the languages to be challenging for translation; a secondary consideration here had to

do with the availability of potential participants within the university for eye tracking purposes.

For the satisfaction experiments performed via the web survey, the same 140 articles selected for the EN_Source experiment were also selected for the translated content experiments, with their PEz and PEp version, and published online.

For the usability, quality and satisfaction (post-task satisfaction questionnaire and moderators' ratings), the same six articles selected for the EN_Source were used in their PEz and PEp versions.

The baseline text used for the translated content experiment was the respective translations of the EN_Source (see 4.2.2.1).

4.3.2.2 Tools

4.3.2.2.1 Spreadsheet Software

Similar to the EN_Source experiments, the language experiments make use of the Spreadsheet Software described in Section 4.2.2.1, with the exception that instead of English, the language displayed in the user interface was that of each participants', that is, German participants would see the interface in German, Chinese participants would see the interface in Simplified Chinese and Japanese participants would see the user interface in Japanese.

4.3.2.2.2 Eye Tracker Device

The eye tracker used in the usability experiments for German, Simplified Chinese and Japanese is the same as that described in Section 4.2.2.2.2.

4.3.2.2.3 Web Survey

The web survey displayed in the webpage was the same for the English language experiments (see 4.2.2.2.5). The survey question *"Was this information helpful?"* (YES/NO) was displayed in German, Simplified Chinese and Japanese on their respective webpages

4.3.2.2.4 Post-task Satisfaction Questionnaire

As described previously in Section 4.2.2.2.6, a post-task questionnaire was presented to all participants after the performance of the usability experiments. All participants saw the questionnaire in English. Nine questions were presented with a Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree).

4.3.2.2.5 Translation Quality Assessment

Quality assessment in the localisation industry (to which the content type belongs – see Section 4.3.2.1) is normally measured using error typologies. As discussed in Chapter 2, these typologies are often developed and customised from the LISA QA metric (see O'Brien 2012, O'Brien et al. 2011, Drugan 2013, etc.) This QA exercise is normally carried out by linguists/translators/moderators in language service providers (LSPs). To appeal to ecological validity for the methods implemented, moderators in an LSP were asked to fill in a questionnaire pertaining to the quality of content for the PEz and PEp versions, for all three languages. The TQA questionnaire used in this research is a tailored version from the freely available KantanMT's framework.[23] When designing the TQA questionnaire, some of the points the industry partner is concerned about were taken into consideration. As a final result, the TQA questionnaire consists of four error categories: adequacy, fluency, syntax and grammar (spelling and sentence structure), and style (terminology and country standards). The questionnaire also presented one satisfaction question.

For adequacy and fluency, a 1-4 Likert scale was used, whereas for syntax and grammar, and satisfaction a 1-3 Likert scale was used (see Section 4.4.1.3 for detailed description of the measures).

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

4.3.3 Procedure

4.3.3.1 Translation and Post-editing

The MT system used 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 post-editing of the MT output was performed by the company's LSP using the guidelines developed by the company. Differently from full post-editing, which the industry partner considers to be the same as HT since the quality has to be the same, light post-editing 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. The guideline consisted of twelve main points:

1. The translation should be semantically correct.
2. Ensure that no information has been accidentally added or omitted.
3. Edit any offensive, inappropriate or culturally unacceptable content.
4. Use as much of the raw MT output as possible.
5. Basic rules regarding spelling apply.
6. No need to implement corrections that are of a stylistic nature only.
7. No need to restructure sentences solely to improve the natural flow of the text.
8. Make sure that terminology is accurate, and that text is not translated if not needed (ex., error messages)
9. Any technical terms - words or phrases with a technical/domain-specific meaning in the source sentence - are recognized and translated accurately.
10. The translation is fully understandable for the intended target user in the target language and conveys the same meaning (propositional content) as the source sentence.
11. Make sure the product names always stay in English! It is a known bug of the MT engine that sometimes it translates the product names.
12. A technically accurate translation should also conform to target language conventions - in terminology, company specific or country-general.

4.3.3.2 Usability Experiments

4.3.3.2.1 Recruitment survey

The recruitment survey was the same described in 4.2.3.1.1.

4.3.3.2.2 Tasks

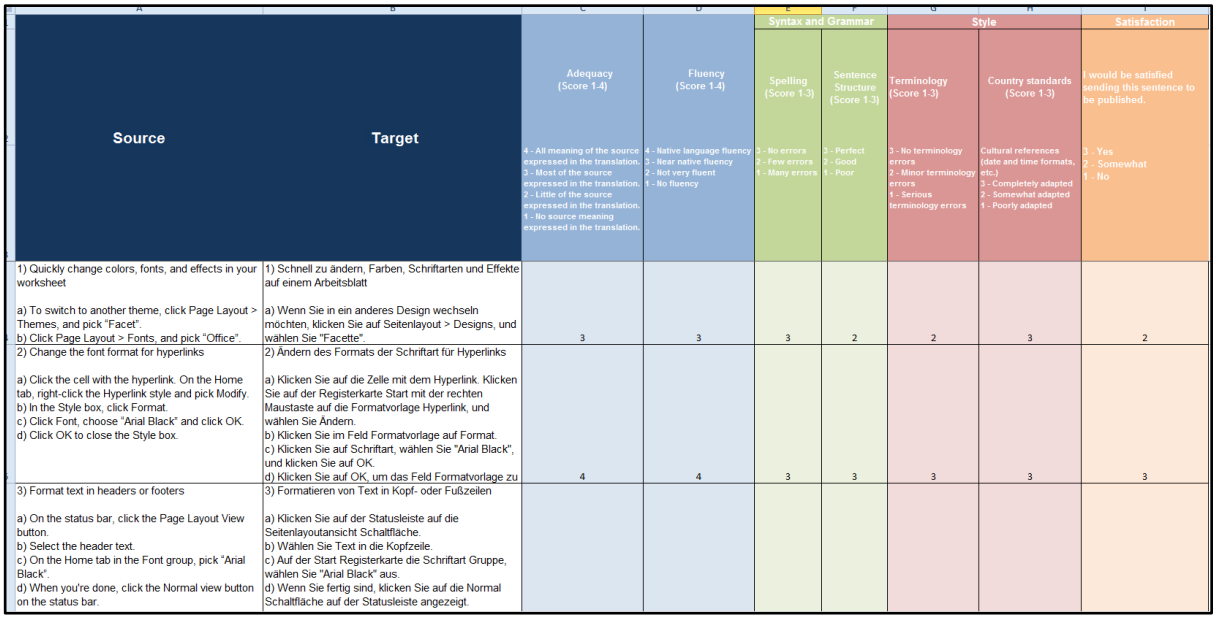
The procedures for the performance of the tasks were the same as the ones described in 4.2.3.1.2, with the exception that for German, Japanese and Simplified language experiments, participants were divided into two groups; one group used the raw machine translated (PEz) instructions and the other used the post-edited (PEp) instructions. Neither group knew that the texts they were reading had been translated as we did not want the participants to be biased (see Figure 4:2 for task design).

4.3.3.3 Quality Experiments

The TQA questionnaire described in Section 4.3.2.2.5 was distributed to eighteen moderators from the company's LSP in February 2016 in a spreadsheet. The content used for the TQA is the same used to perform the usability experiments (described in 4.3.3.2). It was agreed with the industry partner that instead of sentences the moderators would be rating the topic, that is, each task that was presented in the usability experiment. This method was chosen because it is more representative of the customer experience, since the customer is interested if the topic is overall helpful even if some parts of it are unsatisfactorily translated. Both the source and the translation (MT and PE) were shown for the moderator. Figure 4:3 shows the set up for the TQA questionnaire.

Two different types of files were prepared for each language: type A and type B. This was done so the different versions (MT or PE) could be alternated, for example, type A files would have the MT version of task 1, while type B would have the PE version of task 1 – and so on. The two human translated tasks were also rated by the moderators, and the order of the tasks was kept in the original, that is,

as it was presented in the usability experiment. Table 4:1 illustrates how the topics were alternated.



Source	Target	Adequacy (Score 1-4) 4 - All meaning of the source expressed in the translation. 3 - Most of the source expressed in the translation. 2 - Little of the source expressed in the translation. 1 - No source meaning expressed in the translation.	Fluency (Score 1-4) 4 - Native language fluency 3 - Near native fluency 2 - Not very fluent 1 - No fluency	Syntax and Grammar: Spelling (Score 1-3) 3 - No errors 2 - Few errors 1 - Many errors	Syntax and Grammar: Sentence Structure (Score 1-3) 3 - Perfect 2 - Good 1 - Poor	Style: Terminology (Score 1-3) 3 - No terminology errors 2 - Minor terminology errors 1 - Serious terminology errors	Style: Country standards (Score 1-3) Cultural references (date and time formats, etc.) 3 - Completely adapted 2 - Somewhat adapted 1 - Poorly adapted	Satisfaction I would be satisfied sending this sentence to be published. 3 - Yes 2 - Somewhat 1 - No
1) Quickly change colors, fonts, and effects in your worksheet a) To switch to another theme, click Page Layout > Themes, and pick "Facet". b) Click Page Layout > Fonts, and pick "Office".	1) Schnell zu ändern, Farben, Schriftarten und Effekte auf einem Arbeitsblatt a) Wenn Sie in ein anderes Design wechseln möchten, klicken Sie auf Seitenlayout > Designs, und wählen Sie "Facette".	3	3	3	2	2	3	2
2) Change the font format for hyperlinks a) Click the cell with the hyperlink. On the Home tab, right-click the Hyperlink style and pick Modify. b) In the Style box, click Format. c) Click Font, choose "Arial Black" and click OK. d) Click OK to close the Style box.	2) Ändern des Formats der Schriftart für Hyperlinks a) Klicken Sie auf die Zelle mit dem Hyperlink. Klicken Sie auf der Registerkarte Start mit der rechten Maustaste auf die Formatvorlage Hyperlink, und wählen Sie Ändern. b) Klicken Sie im Feld Formatvorlage auf Format. c) Klicken Sie auf Schriftart, wählen Sie "Arial Black", und klicken Sie auf OK. d) Klicken Sie auf OK, um das Feld Formatvorlage zu	4	4	3	3	3	3	3
3) Format text in headers or footers a) On the status bar, click the Page Layout View button. b) Select the header text. c) On the Home tab in the Font group, pick "Arial Black". d) When you're done, click the Normal view button on the status bar.	3) Formatieren von Text in Kopf- oder Fußzeilen a) Klicken Sie auf der Statusleiste auf die Seitenlayoutansicht Schaltfläche. b) Wählen Sie Text in die Kopfzeile. c) Auf der Start Registerkarte die Schriftart Gruppe, wählen Sie "Arial Black" aus. d) Wenn Sie fertig sind, klicken Sie auf die Normal Schaltfläche auf der Statusleiste angezeigt.							

Figure 4:3 - TQA Questionnaire

Task	1	2	3	4	5	6	7	8
Type A	MT	PE	MT	HT	PE	MT	PE	HT
Type B	PE	MT	PE	HT	MT	PE	MT	HT

Table 4:1 - Distribution of Topic for the TQA Questionnaire

The moderators were asked to look at the source and the translated segments and judge them according to the six categories. It was ensured that the moderators were not aware that the files they were assessing contained raw machine translated, lightly post-edited and human translated segments. After rating each segment, the moderators were asked to answer if they were satisfied with the translation. In total, three moderators (for each language) saw the type A file, and three saw the type B file.

4.3.3.4 Satisfaction Experiments

4.3.3.4.1 Web Survey

As mentioned in sections 4.3.2.2.3, a web survey was displayed on the industry partner's website to gather information on 'how useful' the content is for the end user. The articles were published on the company's website first in their PEz version and afterwards in their PEp version.

The articles were published online at different times, depending on several factors, such as:

- When the MT versions were delivered by the LSP and were ready to be published
- When the MT versions gathered a sufficient number of ratings
- When the PE versions were delivered by the LSP and were ready to be published
- When the PE versions gathered a sufficient number of ratings

At this point it is important to mention that some languages tend to get a higher number of ratings than other languages, for example, articles in English tend to get more ratings than articles in Chinese. Therefore, to approximate the number of ratings for the PEz and PEp versions, different periods had to be chosen. Table 4:2 illustrates the time period:

PEz	DE	ZH	JP
Period	JUL-SEPT (3)	JUN-OCT (4)	APR-SEPT (6)
Rating Count	247	98	151
PEp	DE	ZH	JP
Period	OCT-DEC (3)	NOV-JAN (3)	NOV-JAN (3)
Rating Count	438	92	233

Table 4:2 - Period for Web Survey per Language

4.3.3.4.2 Post-task satisfaction Questionnaire

The post-task questionnaire (described in 4.2.2.2.6), designed to capture the users' satisfaction level after performing specific tasks, was presented for all the

participants after they had performed the usability experiment with the PEz and PEp instructions.

4.3.3.4.3 Moderators' ratings (TQA)

The distribution of the TQA questionnaire was described in Section 4.3.3.3.

4.4 Measures

As stated previously, this research focuses on acceptability as a wider model in which usability, satisfaction and quality are elements. This section will describe how each element of acceptability is assessed regarding metrics and measures for all the languages. Table 4:3 summarises the usability, quality and satisfaction experiments as well as the measures used for the English language, while Table 4:4 shows this for the German language, Table 4:5 Simplified Chinese and Table 4:6 for Japanese.

<table>
<tr><th colspan="6">EN</th></tr>
<tr><th rowspan="2">Measures</th><th rowspan="2">Participants</th><th colspan="2">Materials</th><th colspan="2" rowspan="2">Measures</th></tr>
<tr><th>Content</th><th>Tools</th></tr>
<tr><td rowspan="3">Usability</td><td>8 English native speakers</td><td rowspan="9">6 Online Help Content articles (8 tasks)</td><td rowspan="3">Eye-tracking</td><td>Effectiveness</td><td>goal completion</td></tr>
<tr><td>Efficiency</td><td>task time and goal completion</td></tr>
<tr><td>7 English native speakers*</td><td>Cognitive Effort</td><td>fixation duration
fixation count
visit duration
visit count</td></tr>
<tr><td rowspan="3">Quality</td><td rowspan="3">N/A</td><td rowspan="2">Source Content Profiler</td><td>SCP Score</td><td>N° of grammar issues
N° of spelling issues
N° of passive voice issues
Percentage of sentences with unusual POS sequences
Average sentence length
Average word length</td></tr>
<tr><td colspan="2">Domain Classification</td></tr>
<tr><td>Coh-Metrix</td><td colspan="2">Text Easability
Syntactic Simplicity
Referential Cohesion
Verb Cohesion
Readability (Flesch reading ease)</td></tr>
<tr><td rowspan="2">Satisfaction</td><td>8 English native speakers</td><td>Post-Task Satisfaction Questionnaire</td><td>User ratings for satisfaction</td><td>Likert scale</td></tr>
<tr><td>Ratings from real end users of company’s website</td><td>140 Online Help articles**</td><td>Web Survey</td><td>Web user ratings</td><td>"Was this information useful?” YES/NO</td></tr>
</table>

Table 4:3 - Experiments and measures for Source Content

* For the cognitive effort measures, we only use recordings that reach 80% or higher in quality. Therefore, one recording was excluded from this assessment.

**Including the six articles used for the usability experiments, from which eight tasks were created.

German					
Measures	**Participants**	**Materials**		**Metrics**	
		Content	**Tools**		
Usability	14 German native speakers: 8 - DE_PEz 6 - DE_PEp	6 Online Help Content articles (MT and PE) (8 tasks)	Eye-tracking	Effectiveness	goal completion
				Efficiency	Task time and goal completion
	13 German native speakers*: 7 - DE_PEz 6 - DE_PEp			Cognitive Effort	fixation duration fixation count visit duration visit count
Quality	6 professional moderators		TQA questionnaire	Adequacy Fluency Spelling Sentence structure Terminology Country standards	Likert scale
Satisfaction	14 German native speakers: 8 - DE_PEz 6 - DE_PEp		Post-Task Satisfaction Questionnaire	User ratings for satisfaction	Likert scale
	6 professional moderators		TQA questionnaire	Satisfaction	Likert scale
	ratings from real end users of company's website	140 Online Help articles machine translated**	Web Survey	Web user ratings	"Was this information useful?" YES/NO

Table 4:4 - Experiments and measures for Translated Content - German

* For the cognitive effort measures, we only use recordings that reach 80% or higher in quality. Therefore, one recording was excluded from this assessment.

**Including the six articles used for the usability experiments, from which eight tasks were created.

Simplified Chinese					
Measures	**Participants**	**Materials**		**Metrics**	
		Content	**Tools**		
Usability	21 Chinese native speakers: 11 - ZH_PEz 10 - ZH_PEp	6 Online Help Content articles (MT and PE) (8 tasks)	Eye-tracking	Effectiveness	goal completion
				Efficiency	Task time and goal completion
	15 Chinese native speakers*: 7 - ZH_PEz 8 - ZH_PEp			Cognitive Effort	fixation duration fixation count visit duration visit count
Quality	6 professional moderators		TQA questionnaire	Adequacy Fluency Spelling Sentence structure Terminology Country standards	Likert scale
Satisfaction	21 Chinese native speakers: 11 - ZH_PEz 10 - ZH_PEp		Post-Task Satisfaction Questionnaire	User ratings for satisfaction	Likert scale
	6 professional moderators		TQA questionnaire	Satisfaction	Likert scale
	ratings from real end users of company's website	140 Online Help articles machine translated**	Web Survey	Web user ratings	"Was this information useful?" YES/NO

Table 4:5 - Experiments and measures for Translated Content - Simplified Chinese

* For the cognitive effort measures, we only use recordings that reach 80% or higher in quality. Therefore, six recordings were excluded from this assessment.

**Including the six articles used for the usability experiments, from which eight tasks were created.

<table>
<tr><th colspan="6">Japanese</th></tr>
<tr><th rowspan="2">Measures</th><th rowspan="2">Participants</th><th colspan="2">Materials</th><th colspan="2" rowspan="2">Metrics</th></tr>
<tr><th>Content</th><th>Tools</th></tr>
<tr><td rowspan="3">Usability</td><td>28 Japanese native speakers:

13 - JP_PEz
15 - JP_PEp</td><td rowspan="6">6 Online Help Content articles (MT and PE) (8 tasks)</td><td rowspan="3">Eye-tracking</td><td>Effectiveness</td><td>goal completion</td></tr>
<tr><td>14 Japanese native speakers*:

7 - JP_PEz
7 - JP_PEp</td><td>Efficiency</td><td>Task time and goal completion</td></tr>
<tr><td>Cognitive Effort</td><td>fixation duration
fixation count
visit duration
visit count</td></tr>
<tr><td>Quality</td><td>6 professional moderators</td><td>TQA questionnaire</td><td>Adequacy
Fluency
Spelling
Sentence structure
Terminology
Country standards</td><td>Likert scale</td></tr>
<tr><td rowspan="3">Satisfaction</td><td>28 Japanese native speakers:

13 - JP_PEz
15 - JP_PEp</td><td>Post-Task Satisfaction Questionnaire</td><td>User ratings for satisfaction</td><td>Likert scale</td></tr>
<tr><td>6 professional moderators</td><td>TQA questionnaire</td><td>Satisfaction</td><td>Likert scale</td></tr>
<tr><td>ratings from real end user of company's website</td><td>140 Online Help articles machine translated**</td><td>Web Survey</td><td>Web user ratings</td><td>"Was this information useful?" YES/NO</td></tr>
</table>

Table 4:6 - Experiments and measures for Translated Content - Japanese

* For the cognitive effort measures, we only use recordings that reach 80% or higher in quality. Therefore, fourteen recordings were excluded from this assessment.

**Including the six articles used for the usability experiments, from which eight tasks were created.

4.4.1 Usability

Effectiveness

Effectiveness is measured through goal completion, that is, how successful the users were at accomplishing tasks documented in the instructions measured by observing the user interactions as recorded by an eye tracker.

Efficiency

Efficiency is measured via i) total task time, and as ii) the number of successful tasks completed (out of all possible tasks) when total task time is taken into account (Doherty and O'Brien 2014):

$$\sum \frac{accuracy}{total\_task\_time\,(sec.)} x\,100, \quad where \quad \frac{task\_sucesses}{total\_tasks} x\,100 = accuracy$$

A third measure of efficiency is cognitive effort, i.e. how much cognitive effort is evident when users are reading the instructions and trying to complete their tasks? The cognitive effort is measured using the eye tracking metrics of visit duration (seconds), visit count, fixation duration (seconds), and fixation count:

Fixation Duration

Fixation Duration (FD) is the length of fixations for all the fixations within an area of interest (AOI). The longer the fixations are, the higher the cognitive effort is deemed to be.

Fixation Count

Fixation Count (FC) is the total number of fixations within an AOI. High amount of fixations indicates high cognitive effort required.

Visit Duration

Visit duration (VD) is the total time (in seconds) spent looking at an AOI, starting with a fixation within the AOI and ending with a fixation outside this AOI, that is, saccades (or rapid eye movements between fixations) are also counted.

Visit Count

Visit Count (shifts of attention) is the number of visits (using eye movements as evidence) to an AOI. Multiple shifts of attention imply a cost in terms of cognitive effort.

4.4.2 Quality

4.4.2.1 Source Content

Evaluating the quality of the source content, EN_Source, is an important part of this research project. We aim at investigating how readable and accurate the source content is; that is, whether the source content contains any errors (e.g. grammar, syntax etc.) or misinformation, as well as its level of complexity—all of which, if unidentified, could be automatically transferred to the translated text. In order to assess that, we use Source Content Profiler tool (see 4.2.2.2.3) – to assess the features of the content, and Coh-Metrix software (see 4.2.2.2.4) – for text easability, and readability scores.

Source Content Profiler

SCP Score

This measure reflects the quality of a document on a scale from 0 to 100, with a lower score indicating higher quality of the document. It is calculated according to sub-scores such as:

- Number of grammar issues;
- Number of spelling issues;
- Number of passive voice issues
- Percentage of sentences with unusual POS sequences
- Average sentence length
- Average word length

Domain Classification

The Source Content Profiler tool also displays the percentage of to what domain the text belongs. The domain detection feature is based on a machine learning approach based on the domain in which the tool is trained.

Coh-Metrix

Text Easability

Text easability is a Coh-Metrix measure which provides metrics of text characteristics on multiple levels of language and discourse (McNamara et al. 2014). The scores are displayed in percentile, where higher scores mean the text is likely to be easier to read. For this experiment, three components are used:

Syntactic Simplicity – indicates the extent to which sentences in the text use simpler syntactic structures and have fewer words, which makes the text less challenging to read.

Referential Cohesion – indicates whether the text contains words and ideas that overlap across sentences and entire text, that is, has a high referential cohesion.

Verb Cohesion – indicates the extent to which the text contains overlapping verbs. The text is likely to include more coherent event structures when verbs are repeated.

Readability

Readability is measured via the traditional Flesch reading ease measure (Flesch, 1948).

Flesch Reading Ease – The output is a number from 0 to 100, where a higher score indicates the text is easier to read.

4.4.2.2 Translated Content

As described in 4.3.2.2.5, the TQA questionnaire consisted of four categories: adequacy, fluency, syntax and grammar, and style (satisfaction measure is described below in Section 4.3.3.4.3 - satisfaction) and was used to evaluate the PEz version of the articles for all the three languages – DE, ZH and JP.

Adequacy and Fluency are measured with a 1-4 Likert scale:

Adequacy

4 – All meaning expressed in the source fragment appears in the translation fragment.

3 – Most of the source fragment meaning is expressed in the translation fragment.

2 – Little of the source fragment meaning is expressed in the translation fragment.

1 – None of the meaning expressed in the source fragment is expressed in the translation fragment.

Fluency

4 – Native language fluency. No grammar errors, good word choice and syntactic structure. No post-editing required.

3 – Near native fluency. Few terminology or grammar errors which don't impact the overall understanding of the meaning. Little post-editing required.

2 – Not very fluent. About half of translation contains errors and requires post-editing.

1 – No fluency. Absolutely ungrammatical and for the most part doesn't make any sense. Translation has to be re-written from scratch.

Syntax and Grammar are measured in a Likert scale from 1-3, via spelling and sentence structure:

Spelling

3- No spelling errors

2- Few spelling errors

1- Many spelling errors

Sentence Structure

3- perfect

2- good

1- poor

Style is measure via a 1-3 Likert scale via Terminology and country standards:

Terminology

3- No terminology errors

2- Minor terminology errors, normally associated with differences on gender, number, preposition, article, or verb tense that meet all the 3 conditions below:

- Do not modify or misrepresent the functionality of the product
- Do not appear in an important or highly visible location such as the menu bar or a command
- Correction has not been previously requested by LQA LS (Language Service – Language Quality Assessment)

1- Terminology errors are misleading to the users, jeopardizing the comprehension of the text of misrepresenting the functionality of the product.

Country standards

The translation must correctly adapt cultural references (date and time formats, units of measurement, currency, number formats, sorting order etc.)

3- Completely adapts to country standards

2- Somewhat adapts to country standards

1- Poorly adapts to country standards

4.4.3 Satisfaction

The web survey gathered a Boolean answer from real end users of the online articles. The survey question “Was this information helpful” had a simple YES/NO answer. The post-task satisfaction questionnaire contained 9 questions with multiple choice answers in a Likert scale from 1-5. For the rating for Satisfaction by the moderators, satisfaction is measured in a 1-3 Likert scale:

I would be satisfied sending this sentence to be published

3- Yes

2- Somewhat

1- No

4.5 Statistical Analysis

For the statistical analysis of the measures presented in Section 4.4, a number of different tests are used, as explained in this section.

ANOVA

Analysis of variance (ANOVA) is a statistical method used to compare the differences among group means (in two or more groups) on a single independent variable (or factor). When one single factor is involved, it is called a one-way ANOVA.

A two-way ANOVA (also called Factorial ANOVA) is an extension of the one-way ANOVA and it analyses the influence of two factors on a single dependent variable. Besides comparing the main effect of each factor on the independent variable, it also analyses the extent to which the two factors may combine to influence scores on the dependent variable (Howitt and Cramer 2005, p.220). In this study, a two-way ANOVA is used, for example, to compare the effect of Post-Editing and Language on Effectiveness.

MANOVA

Multivariate analysis of variance (MANOVA) is an ANOVA with two or more dependent variables. MANOVA combines the dependent variables to see whether the different groups differ in their mean in this combined set of dependent variables (Howitt and Cramer 2011, p.317).

A two-way MANOVA (or Factorial MANOVA) is an extension of the one-way MANOVA and it analyses the influence of two factors on the two (or more) dependent variables. When calculating Factorial MANOVA (and ANOVA) in SPSS[24], 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 in proportion to its sample size is taken into consideration. The pairwise comparisons tables display the factors and dependent variables combined in different ways, so that different interaction can be analysed. In this study, a two-way MANOVA is used, for example, to compare the effect of Post-Editing and Language on fixation count for two different areas of interest (instruction and user interface).

Repeated Measures

Repeated measures designs have the same participants measured in all conditions, that is, it compares the differences in mean scores under two or more different conditions (Howitt and Cramer 2011, p.230). In this study, repeated measures design is used to assess both dependent variables as one. For example, a two-way MANOVA with repeated measures is used to compare the effect of Post-editing and language on fixation count (two different areas of interest - AOI) as well as the effect of fixation duration in the AOI instruction on the AOI user interface and vice-versa.

[24] Statistical Package for the Social Sciences by IBM. See www.ibm.com/analytics/us/en/technology/spss [Last accessed 08 May 2016]

Post Hoc Test

Post-hoc tests are multiple comparisons which determine which conditions differ significantly from each other (Howitt and Cramer 2005, p.250). In the case of this study, a post-hoc is used when one-way ANOVA reveals a significant result, since a pairwise comparison with one factor and one independent variable cannot show where the difference (if any) is.

Significance Level

Generally, researchers report a 0.05 level of significance, which means that the results are not likely to happen by chance more than 5 times in 100 tries – i.e. 95% confidence interval. However, due to the exploratory nature of this research, the cut off for significance used is 0.10 % (Bernard 2011, p.485), which means that the confidence interval is 90%.[25]

4.6 Conclusion

This chapter presented the methodology applied in this research regarding the source content and translated content. The measures for each experiment was described as well as the statistical test applied in order to determine the statistically significant results. The following chapters describe the results for each of the experiments: usability (Chapter 5), quality and satisfaction (Chapter 6).

[25] See http://www.measuringu.com/blog/confidence-levels.php [Last accessed 05 May 2016]

Chapter 5 – Results I

As seen in Chapter 4, the present research applies different and complementary experiments in order to assess the elements of Acceptability: usability, quality and satisfaction. This chapter presents the results for the usability experiments: effectiveness, efficiency and cognitive data. Results for the quality and satisfaction are presented in Chapter 6. The terminology used for all the results (Chapter 5 and 6) are clarified here.

Clarifying Terminology Used Throughout the Results Chapters

The Online Help articles used for the satisfaction experiment via the web survey were machine translated and post-edited, and posted online on the company's website. For the usability, quality and satisfaction (post-task questionnaire[26] and moderators' ratings) experiments, five of those articles were used as instructions for six tasks. One human translated article was selected in order to be used as instructions for two tasks which were added as control tasks. In total, eight tasks were created: six tasks used the machine translated or post-edited instructions, and two tasks used the human translated instructions. The source text of the instructions was also used by the English participants to perform the tasks in the usability experiments, and post-task satisfaction questionnaire. Table 5:1 illustrates the task design.

	Instruction Type							
Groups	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8
PEz (No PE)	MT	MT	MT	HT	MT	MT	MT	HT
PEp (Light PE)	PE	PE	PE	HT	PE	PE	PE	HT
EN_Source	source	source	source	source	source	source	source	source

Table 5:1 - Task Design

For the analysis of the results, the instructions and articles were divided into two instruction types:

[26] Displayed after the usability experiments

1) Instructions which were machine translated and post-edited (tasks 1, 2, 3, 5, 6 and 7), that is, the machine translated and lightly post-edited instructions – hereafter "MT Instructions", which includes the PE_Levels PEz (zero post-editing), PEp (light post-editing), and also the source of those translations (used by the English participants);

2) Instructions which were human translated (task 4 and 8), that is, the human translated instructions – hereafter "HT Instructions" included as control tasks into the PEz and PEp instructions, and also the source of those translations (used by the English participants);

Table 5:2 illustrates this division per instruction type:

	Instruction Type							
Groups	**MT Instructions**						**HT Instructions**	
PEz (No PE)	Task 1	Task 2	Task 3	Task 5	Task 6	Task 7	Task 4	Task 8
PEp (Light PE)	Task 1	Task 2	Task 3	Task 5	Task 6	Task 7	Task 4	Task 8
EN_Source (source)	Task 1	Task 2	Task 3	Task 5	Task 6	Task 7	Task 4	Task 8

Table 5:2 - Division per instruction type

Therefore, "MT Instructions" refers to both PEz and PEp instructions for tasks 1,2,3,5,6 and 7 – and the English source of those; while HT Instructions refers to the instructions which were human translated and incorporated into the PEp and PEz instructions for tasks 4 and 8 – and the English source of those. It is important to note that throughout this chapter, when results for the MT Instructions are reported, the results for tasks 4 and 8 are not included because they involved human translation, and vice-versa.

For all of the results presented below, "Language" refers to the languages investigated:

i. English (EN);
ii. German (DE);
iii. Simplified Chinese (ZH);
iv. Japanese (JP).

However, the term “translated content” is used to refer exclusively to the target languages DE, ZH and JP, whereas “source content” refers to the EN language.

The term “PE_LEVEL” represents the level of post-editing applied to the task instructions:

i. “PEp” refers to the professional light post-editing implemented in the instructions;
ii. “PEz” (post-editing zero) refers to the raw-machine translated instructions;
iii. “Source” refers to the English source instructions.

Note that PEp, PEz and Source may also refer to the **groups** that used those specific instructions types (see Table 5:1 and Table 5:2), that is, ‘PEz groups’ means all groups that used raw machine translated instructions, including all languages. When analysing specific **groups**, a joint terminology is used:

i. DE_PEz refers to the German participants who used the raw machine translated instructions (of the MT Instructions) to perform the tasks 1, 2, 3, 5, 6 and 7, along with the two HT tasks (4 and 8). Therefore, when comparison is made for the HT instructions, DE_PEz refers to the group that performed tasks 4 and 8 (HT instructions) along with the raw machine translation instructions.
ii. DE_PEp refers to the German participants who used the lightly post-edited instructions (of the MT Instructions) to perform the tasks 1, 2, 3, 5, 6 and 7, along with the two HT tasks (4 and 8). Therefore, when comparison is made for the HT instructions, DE_PEp refers to the group that performed tasks 4 and 8 (HT instructions) along with the lightly post-edited instructions. And so on for ZH (ZH_PEz and ZH_PEp) and JP (JP_PEz and JP_PEp).
iii. EN_Source refers to the participants who used the source English text instructions to perform the tasks (task 1, 2, 3, 5, 6 and 7 – and also tasks 4 and 8).

In order to report the results clearly, the chapter is organized according to the experiments that were implemented: usability (5.1), quality (Chapter 6, Section 6.1) and satisfaction (Chapter 6, Section 6.2). For each experiment, results of the MT instructions for the translated content (DE, ZH and JP) are presented first (not including results for the HT instruction), followed by a comparison with the English source instructions/group. Subsequently, results for the HT instructions for the translated content are presented (not including results for the MT instruction tasks) also followed by a comparison with the English source instructions/group.

5.1 Usability

As discussed in Chapter 3, Section 3.1.1, usability is operationalised using the ISO definition and is composed of three concepts: efficiency, effectiveness and satisfaction. Section 5.1.1 describes the participants' background regarding English proficiency level, version of the spreadsheet application they have used, and the frequency they use spreadsheet application (any version). It is worth mentioning that the participants are the same who performed the post-task satisfaction questionnaire (described in Chapter 6, Section 6.2.1).

5.1.1 Participants Background

In the pre-task survey, the participants were asked about their age, gender, education level (see Chapter 4, Section 4.2.1.1 and 4.3.1.1). A question about English proficiency level was presented and options from the Common European Framework of Reference for Languages[27] were given along with the explanation for each. Question about the participants' knowledge on office tools were also presented. The questions were:

- What is your proficiency level in English?
- Have you ever used office tools?

[27] http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp [Last accessed 11 May 2016]

- Specify the versions [of office tools] you have used.
- How often do you use office tools?

Table 5:3 shows the results for the English language proficiency, Table 5:4 shows results for software version, while Table 5:5 shows the results for frequency of use.

English Proficiency	DE		ZH		JP	
	PEz	PEp	PEz	PEp	PEz	PEp
C2 Proficiency	3	2	1	1	1	0
C1 Advanced	4	3	2	4	1	2
B2 Upper Intermediate	1	1	5	3	5	7
B1 Intermediate	0	0	3	1	5	5
A2 Elementary	0	0	0	1	0	0
A1 Beginner	0	0	0	0	1	1
total participants	8	6	11	10	13	15

Table 5:3 - English Proficiency

Software version 2013	EN	DE		ZH		JP	
	SOURCE	PEz	PEp	PEz	PEp	PEz	PEp
Yes	3	6	3	5	6	6	8
No	5	2	3	6	4	7	7
total participants	8	8	6	11	10	13	15

Table 5:4 - Usage of Software version 2013

How often do you use spreadsheet applications?	EN	DE		ZH		JP	
	SOURCE	PEz	PEp	PEz	PEp	PEz	PEp
every day	0	0	1	3	1	2	5
one to three times a week	6	5	1	7	4	2	4
once a month	2	1	2	1	5	7	6
never	0	2	2	0	0	2	0
total participants	8	8	6	11	10	13	15

Table 5:5 - Frequency of usage

Regarding the question “Have you ever used office tools?” all participants from all languages answered that they have used office tools before. This indicates that all participants were literate in these types of applications and would not have problems such as understanding what a ‘cell’ or a ‘button’ is.

Regarding the English proficiency, the German participants show the highest levels of proficiency in English, having their answers between C2 – Proficiency level to C1 – Advanced. Only two German participants (one of each group) answered they had a B2 – Upper Intermediate level. For the Simplified Chinese participants, the majority said they between C1 – Advanced level to B1 – Intermediate level. Two participants (one of each group) said their proficiency level was a C2 – Proficiency, and one participant form the PEp groups said she/he had an A2 – Elementary level. Finally, the Japanese languages have the majority of participants between C1 – Advanced to B1 – Intermediate levels, similarly to the Simplified Chinese language. One participant (PEz group) answered she/he had a C2 – Proficiency level, while two participants (one of each group) answered they had an A1 – Beginner level.

When asked to specify the versions of spreadsheet application they have already used, a balanced number of participants who have and have not used the 2013 version can be seen for all languages. When asked how often they use spreadsheet applications (any version), the majority of participants of the German language seemed to use it between one-three times a week, to once a month. For the Simplified Chinese language, the majority of participants answered they used spreadsheet application between every day to once a month. Finally, The Japanese participants also answered they used the application between every day to once month, hover, two participants (PEz) answered they never used the application.

5.1.2 Usability Experiments

The Usability experiments intend to answer the following research questions:

***RQ1**: Does Post-editing level have an effect on usability?*

***RQ4**: How do different target languages compare in terms of usability for both PEp and PEz content?*

***RQ7**: How does usability of Source Content compared with usability of the translated content (PEp and PEz)?*

For the analysis of effectiveness, and efficiency, a Factorial ANOVA (two-way ANOVA) is applied since it not only tests the main effect of every factor (independent variables) but also whether there is any interaction between them (see Section 4.5, Chapter 4). Factorial ANOVA provides three results: factor 1, factor 2, and the interactions between these factors (factor1*factor2). Therefore, the analysis presented starts with a two-way ANOVA for the languages (factor 1) and their PE_LEVELs (factor 2), which aims to establish relationships between Languages, PE_LEVEL and the interactions between the two of them (Language*PE_LEVEL). Pairwise comparison is implemented in order to investigate whether there are differences in the interaction between the factors language and post-editing level (Language*PE_LEVEL), as well as post-editing level and language (PE_LEVEL*Language).

When comparing the source, a one-way ANOVA is also used to compare the English language group against the other languages and post-editing levels, where a Tukey post-hoc is also computed. Post-hoc tests calculate the variances between the subgroups (i.e., EN_Source vs DE_PEp; EN_Source vs DE_PEz).

For the Cognitive data analysis, a two-way MANOVA is implemented because MANOVA allows testing two or more dependent variables that may have a correlation (in this case, fixations and visits) between them. Because of that, repeated measures approach was also implemented in order to assess both dependent variables as one. Similarly to the two-way ANOVA, the two-way MANOVA also provides the three results as described above. When comparing the source, a one-way MANOVA with repeated measures is conducted. Due to the exploratory nature of this research, the cut off for significance used is 0.10, which means that the confidence interval is 90% (Bernard 2011, p.485).[28]

5.1.3 Effectiveness (Goal Completion)

Effectiveness is measured through goal completion (see Chapter 4, Section 4.4.1.1), with the tasks (goals) to be completed by each participant.

[28] See also http://www.measuringu.com/blog/confidence-levels.php [Last accessed 05 May 2016]

5.1.3.1 MT Instructions

A two-way ANOVA was conducted to compare the main effects of the two factors: Language and Post-editing Level (PE_LEVEL) on Effectiveness (goal completion) for the MT instructions. Language consisted of three levels (DE (German), ZH (Simplified Chinese) and JP (Japanese), and PE_LEVEL included two levels (PEz (raw machine translation) and PEp (light professional post-editing).

LANGUAGE: The factor Language was found not to have a statistically significant effect on goal completion (F (2, 57) = 2.29, p>.10). This means that when the factor Language is considered without distinctions between PE_LEVELs, there is no statistically significant difference across the three translated languages, DE (M= 4.13, SE=0.31), ZH (M= 3.28, SE=0.25), and JP (M= 3.49, SE=0.22) for effectiveness.

PE_LEVEL: The factor PE_LEVEL was found to have a very statistically significant effect on goal completion, where F (1, 57) = 14.13, p<.001. This indicates that when the factor PE_LEVEL is considered without distinctions between Languages, there is a statistical difference across the two post-editing levels PEz (M= 3.05, SE=0.21) and PEp (M= 4.21, SE=0.22).

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect on goal completion (F (2, 57) = 1.20, p>.10). This means that the factor language (no distinctions between PE_LEVELs) combined with the factor PE_LEVEL (no distinction between Languages) do not have a joint effect on effectiveness.

Table 5:6 shows the percentage of successfully completed tasks for each post-editing level per language, while Figure 5:1 illustrates the estimated marginal means for each post-editing level (PEz and PEp). As discussed in Chapter 4, estimated marginal means refers to unweighted means, that is, the covariates are held at their mean value (i.e., the mean of PE_LEVEL when Language is held constant at its mean value). This is important when comparing the means of unequal sample sizes where each mean in proportion to its sample size is taken into consideration.

For all languages, participants who used the PEp instructions have a higher percentage of completed tasks when compared to the participants who used the PEz instructions.

GOAL COMPLETION	DE		ZH		JP	
	PEz	PEp	PEz	PEp	PEz	PEp
	53%	84%	47%	61%	51%	64%

Table 5:6 - Goal Completion Percentage - Translated Content

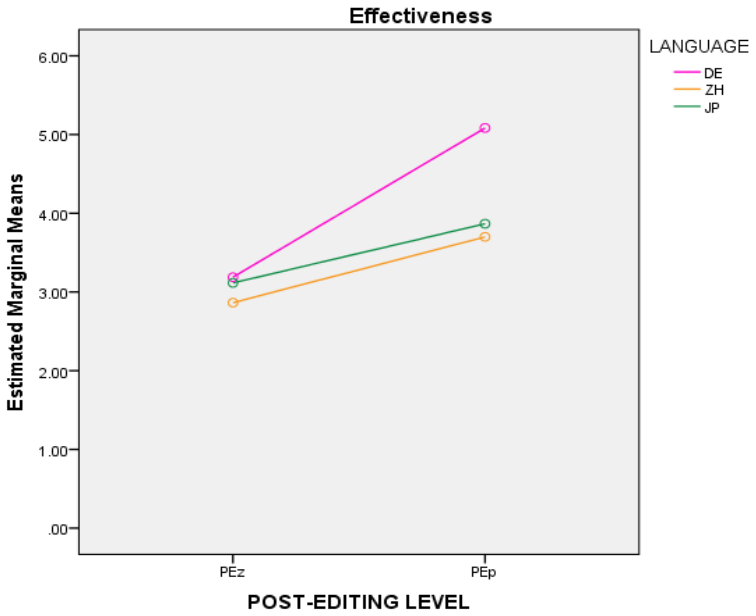


Figure 5:1 - Goal Completion - Translated Content

When looking at Effectiveness across PE_LEVELs (Figure 5:1), the German PEp group shows higher effectiveness when compared to ZH_PEp and JP_PEp groups. This result was statistically significant at the $p<.05$ level. No statistical difference was found between the ZH_PEp and JP_ PEp. These results indicate that German participants who used the lightly post-edited instructions were notably more successful than the Chinese and Japanese participants who also used the lightly post-edited instructions of the language. Regarding the PE_LEVEL PEz, no statistical difference was found across languages ($p>.10$), which demonstrates that all participants who used the raw machine translated instructions (DE_PEZ, ZH_PEZ and JP_PEz) were, in general, as successful when performing the tasks.

When looking at Effectiveness within languages, all PEp groups show higher effectiveness than their PEz groups. In order to identify whether this effect was

statistically significant, a pairwise comparison[29] was computed. Results showed that the DE_PEp (M= 5.08, SE =.47) group presented a very statistically significant difference at the $p<.01$ level against the DE_PEz (M=3.18, SE=.41) group. The Japanese language also showed a statistically significant difference at the $p<.10$ level between its PE_LEVELs JP_PEp (M=3.86, SE =.30) and JP_PEz (M= 3.11, SE= .32), where the PEp group was more successful when compared to the PEz group. The Chinese language also showed a statistically significant difference at the p>.10 level between its PE_LEVELs ZH_PEp (M=3.70, SE=.37) and ZH_PEz (M=2.86, SD=.35).

Overall, participants who used the PEp instructions to perform the tasks were statistically more successful than participants who used the PEz instruction. The German participants who used the PEp instructions were more successful than all the other groups.

Comparison with Source

The performance of the participants who used the English source of the MT instructions was also computed. Table 5:7 shows the percentage of successfully completed tasks for each language and their respective post-editing levels compared to the English source, and Figure 5:2 illustrates the estimated marginal means for each post-editing level compared to the Source.

It is interesting to note that the German participants who used the PEp instructions had the highest number of successfully completed tasks (84%), this being higher than the number of successful tasks completed by the participants who used the English source instructions (65%). However, the difference between DE_PEp (M= 5.05, SD =1.28) was not found to be statistically different ($p>.10$) from the EN_Source (M= 3.93, SD =1.01). The Chinese and Japanese participants who used the PEp versions (ZH_PEp = 61% and JP_PEp = 64%) have a close number of successfully completed tasks when compared to the English (65%). This lack of difference is confirmed by a Tukey post-hoc comparison of the three groups, which found the results for ZH_PEp (M= 3.70, SD = 1.25) and JP_PEp (M= 3.86, SD = .89)

[29] The pairwise comparisons are based on the estimated marginal means.

not to have a statistically significant difference (p>.10). This indicates that light post-editing allowed both JP and ZH groups to succesfully complete around the same amount of tasks as the EN_Source group.

GOAL COMPLETION	EN	DE		ZH		JP	
	SOURCE	PEz	PEp	PEz	PEp	PEz	PEp
	65%	53%	84%	47%	61%	51%	64%

Table 5:7 - Goal Completion Percentage - Source

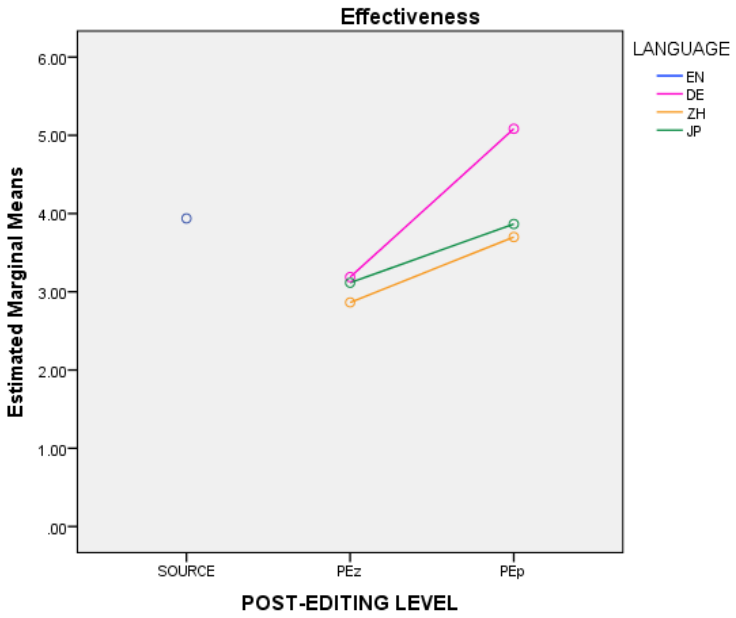


Figure 5:2 - Goal Completion - Source

When looking at the PE_LEVEL PEz in Table 5:7 and Figure 5:2, one may observe that the results for the translated content are lower when compared to the EN_Source. A pairwise comparison showed that there was a statistically significant difference between EN_Source and the PE_LEVEL PEz when all languages are considered (p<.05). However, when looking at multiple comparisons, a Tukey post-hoc did not find any statistically significant difference (p>.10) between EN_Source and DE_PEz (M= 3.18, SD = 1.68); or EN_Source and ZH_PEz (M= 2.86, SD = .95); or EN_Source and JP_PEz (M= 3.11, SD = 1.13). These results indicate that, when compared against each PEz group separately, the difference between the EN_Source is not very significant, even though English participants were able to complete more tasks.

Finally, regarding the comparison between EN_Source and the PE_LEVEL PEp for all languages, no statistically significant difference was found ($p>.10$) which indicates that, in general, participants that used the PE versions of the instructions were as (or more – in the case of the DE_PEp group) successful as the participants using the English source instructions.

Overall, participants who used the PEp instructions were able to successfully complete as many tasks as the EN_Source group.

5.1.3.2 HT Instructions

As stated previously, two sets of instructions were human translated and incorporated into the MT instructions as two control tasks (tasks 4 and 8). Note that this data only involves HT instructions – results for the MT instructions were presented previously in section 5.1.2.1.

A two-way ANOVA was conducted to compare the main effects of Language and PE_LEVEL [30]on goal completion for the human translated tasks.

LANGUAGE: The factor Language was found not to have a statistically significant effect on goal completion (F (2, 57) = 2.12, $p>.10$), which indicates that participants from all languages (without distinctions between PE_LEVELs) were also comparatively successful when performing the tasks with HT instructions – DE (M= 1.87, SE=0.10), ZH (M= 1.69, SE=0.08), JP (M= 1.86, SE=0.07).

PE_LEVEL: The factor PE_LEVEL was not statistically significant, where (F (1, 57) = 2.12, $p>.10$), which indicates that participants from both PEz (M= 1.73, SE=0.07) and PEp (M= 1.88, SE=0.07) groups (without distinctions between Languages) were comparatively successful.

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect on goal completion (F (2, 57) = .433, $p>.10$) for the HT instructions. This means that the factor language (no distinctions between PE_LEVELs) combined with the factor PE_LEVEL (no distinctions between Languages) do not have a joint effect on effectiveness.

[30] PE_LEVEL in this context refers to either the group who used the HT instructions as part of the raw MT instructions or the group who used the HT instructions as part of the lightly post-edited instructions

Table 5:8 shows the percentage of successfully completed tasks for each post-editing level and Figure 5:3 illustrates the estimated marginal means for each group (PEz and PEp).

GOAL COMPLETION HT	DE		ZH		JP	
	PEz	PEp	PEz	PEp	PEz	PEp
	87%	100%	84%	85%	88%	98%

Table 5:8 - Goal Completion Percentage HT Instructions – Translated Content

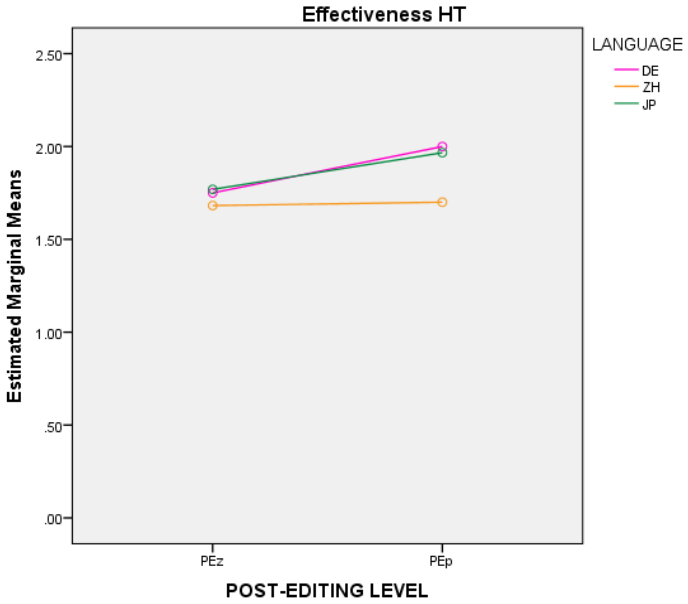


Figure 5:3 - Goal Completion HT Instructions – Translated Content

It is possible to note in Figure 5:3 that the DE_PEp (M= 2.00, SD =0.00), ZH_PEp (M= 1.70, SD =.53) and JP_PEp (M= 1.96, SD =.12) groups have a higher percentage of completed tasks when compared to the participants from the PEz groups (DE_PEz (M= 1.75, SD =.46), ZH_PEz (M= 1.68, SD =.33), JP_PEz (M= 1.76, SD =.56)). However, no statistically significant difference between these groups within languages (p>.10) was found in the pairwise comparisons. These results indicate that all participants from all the PEz groups, although slightly less successful, were, in general, as successful as the participants from the PEp groups when performing the human translated tasks.

Comparison with Source

The performance of the participants who used the English source of the HT instructions was also computed. Table 5:9 shows the percentage of successfully completed tasks for each language and their respective post-editing levels compared to the English source while Figure 5:4 illustrates the estimated marginal means for each post-editing level compared to the English source.

GOAL COMPLETION HT	EN	DE		ZH		JP	
	SOURCE	PEz	PEp	PEz	PEp	PEz	PEp
	100%	87%	100%	84%	85%	88%	98%

Table 5:9 - Goal Completion Percentage HT Instructions - Source

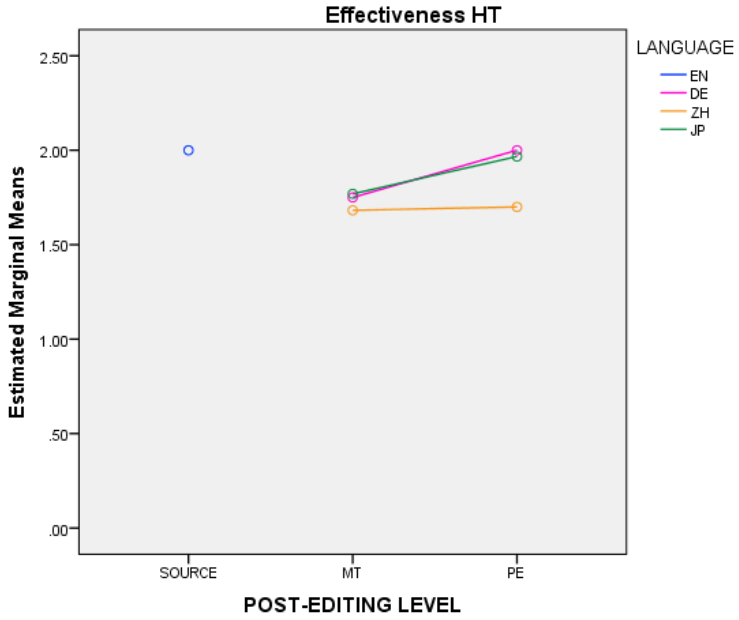


Figure 5:4 - Goal Completion HT Instructions - Source

When compared to German and Japanese languages (no distinctions between PE-LEVELs), the English language was not statistically significant different (p>.10). This indicates that participants from German and Japanese languages (all PE_LEVELs included) were essentially as successful as participants from the EN language. There was a statistically significant difference between the English language (M= 2.0, SD =.0) and the Chinese language (both PE-LEVELs considered) at the p<.05 level. A

Tukey post-hoc, however, did not find statistically significant differences between EN_Source and ZH_PEp or EN_Source and ZH_PEz groups (p>.10).

Overall, these results indicate that all language groups were essentially as successful at performing the tasks with HT instructions when compared to the EN_Source group, which used the source text of the HT instructions.

5.1.4 Results for Efficiency

As seen in Chapter 4, Section 4.4.1.1, efficiency is measured via i) total task time, and as ii) the number of successful tasks completed (out of all possible tasks) when total task time is taken into account. For example, if a participant successfully completed three out of the six tasks, with a mean total task time of 136.20 seconds, it would be calculated as:

$$\sum \left(\frac{3}{6} \; x \; 100 = 50 \right) \frac{50}{136.20} \; x \; 100 = 36.71$$

This section starts by reporting results on total task time, and moving to report results on efficiency (successful tasks/task time).

5.1.4.1 Task Time

5.1.4.1.1 *MT Instructions*

A two-way ANOVA was conducted to compare the main effects of the factors Language and PE_LEVEL on total task time for the MT instructions.

LANGUAGE: The factor Language was found not to have a statistically significant effect on total task time (F (2, 57) = .47, p>.10). This means that when the factor Language is considered without distinctions between PE_LEVELs, there is no statistically significant difference across the three translated languages DE (M=1170.09, SE =93.96), ZH (M=1193.96, SE =76.02) and JP (M=1100.31, SE =65.93) for task time.

PE_LEVEL: The factor PE_LEVEL was found to have a statistically significant effect on task time, where F (1, 57) = 3.47, p<.10. This indicates that when the

factor PE_LEVEL is considered without distinctions between languages, there is a statistically significant difference across the two post-editing levels PEz (M=1240.33, SE=62.76) and PEp (M=1069, SE=66.96).

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect on task time (F (2, 57) = 2.33, p>.10). This means that the factor Language (no distinctions between PE_LEVELs) combined with the factor PE_LEVEL (no distinction between Languages) do not have a joint effect on total task time.

Table 5:10 shows the mean and standard deviation for each language and their respective post-editing levels (in seconds), and Figure 5:5 illustrates the estimated marginal means for each post-editing level.

	Groups	Mean	Std. Deviation
DE	PEz	1225.99	217.55
	PEp	1114.19	307.59
ZH	PEz	1402.12	366.75
	PEp	985.81	380.89
JP	PEz	1092.91	379.59
	PEp	1107.72	350.30

Table 5:10 - Mean and Standard Deviation for Total Task Time (secs) – Translated Content

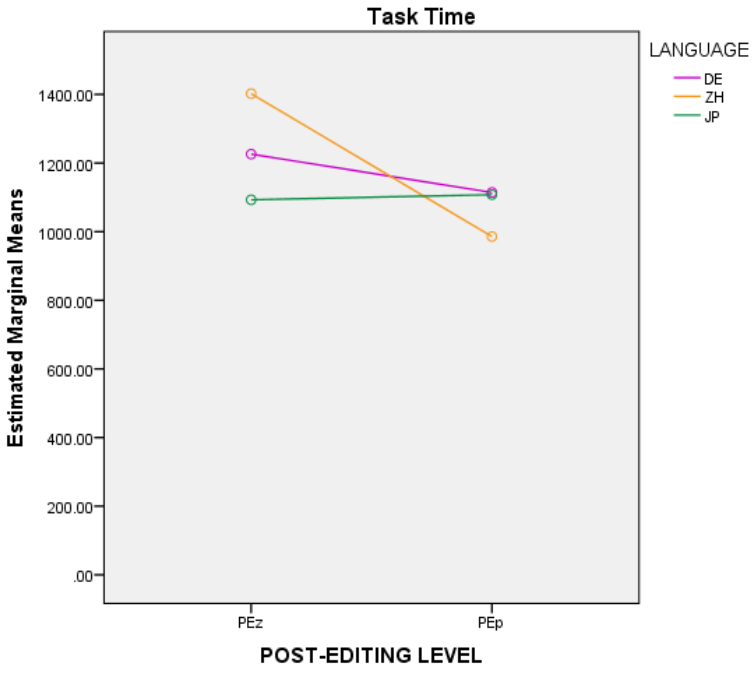


Figure 5:5 - Task Time (secs) – Translated Content

When looking at task time (Figure 5:5) across PE_LEVELs, the ZH_PEp group presents lower times when compared to the DE_PEp and JP_PEp groups, however, there were no statistically significant differences for the PE_LEVEL PEp across the translated languages. This means that participants from all the PEp groups (DE_PEp, ZH_PEp and JP_PEp) were similarly fast when performing the tasks. Regarding the PE_LEVEL PEz, the ZH_PEz presents higher times when compared to the DE_PEz and JP_PEz. This result was a statistically significant only for the ZH_PEz against the JP_PEz group at the $p<.05$ level, which means that the JP_PEz participants were statistically faster when compared against the ZH_PEz group. No statistically significant differences were found between the German and the Japanese PEz groups ($p>.10$) or German and Chinese PEz groups ($p>.10$). These results indicate that participants from the DE_PEz group were as fast as JP_PEz and ZH_PEz groups.

Regarding PE_LEVEL within languages, a pairwise comparison found that the Simplified Chinese language presented a statistically significant difference ZH_PEp and ZH_PEz groups, where $p<.05$. This result indicates that for the Simplified Chinese language, participants who used the raw machine translated instructions (ZH_PEz) were significantly slower than the participants who used the lightly post-edited instructions (ZH_PEp). For the German and Japanese languages, however, no statistically significant difference was found within the PE_LEVELS (DE_PEp and DE_PEz; JP_PEz and JP_PEp), where $p>.10$, even when the PEp group of the Japanese language shows slightly higher task time than the PEz group. The results for German and Japanese indicate that participants who used the PEz instructions were as fast as the participants who used the PEp instructions.

Overall, we can conclude that participants who used the PEz instructions were slower at performing the tasks when compared to participants who used the PEp instructions for the ZH language. For German and Japanese, the time taken by participants to perform the tasks from both groups was comparative.

Comparison with Source

The performance of the participants who used the English source of the MT Instructions was also computed for task time. A pairwise comparison found that the

participants who used the EN instructions were statistically faster at completing tasks when compared to the translated languages (including both PE_LEVELs): DE (M=1170.09, SE=89.41) at the $p<.05$ level; ZH (M=1193.96, SE=72.34) at the $p<.005$ level; and JP (M=1100.31, SE=62.78) at the $p<.05$. There was also a very statistically significant difference between the English source and the PE_LEVEL PEz (M=1240.33, SE=59.72) at the $p<.001$ level, and the PE_LEVEL PEp (M= 1069, SE=63.72) at the $p<0.1$ level.

Table 5:11 shows the mean and standard deviation for each language and their respective post-editing levels (in seconds) compared to the English source, while Figure 5:6 illustrates the estimated marginal means for each post-editing level compared to the Source.

	Groups	Mean	Std. Deviation
EN	SOURCE	780.19	128.31
DE	PEz	1225.99	217.55
	PEp	1114.19	307.59
ZH	PEz	1402.12	366.75
	PEp	985.81	380.89
JP	PEz	1092.91	379.59
	PEp	1107.72	350.30

Table 5:11 - Mean and Standard Deviation for Task Time (secs) – Source

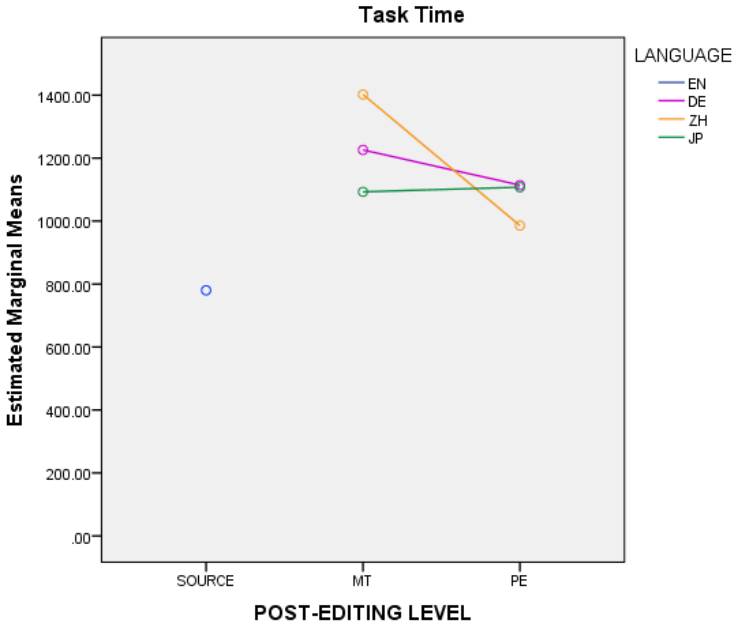


Figure 5:6- Task Time (secs) - Source

Tukey post-hoc results show that participants who used the English source instructions were statistically faster when compared to the participants from the ZH_PEz group (p<.005). No statistically significant difference was found between the EN_SOURCE groups and the other groups (DE_PEz, DE_PEp, ZH_PEp, JP_PEz, and JP_PEp).

5.1.4.1.2 *HT Instructions*

A two-way ANOVA was conducted to compare the main effects of Language and PE_LEVEL on task time for the human translated tasks. Note that this data only involves HT instructions – results for the MT instruction were presented in the previous section (5.1.3.1.1).

LANGUAGE: The factor Language was found not to have a statistically significant effect on total task time for the HT instructions (F (2, 57) = 1.70, p>.10). This means that when the factor Language is considered without distinctions between PE_LEVELs, there is no statistically significant difference across the three translated languages DE (M=188.28, SE =19.09), ZH (M=203.41, SE =15.44) and JP (M=166.16, SE =13.39) for task time.

PE_LEVEL: The factor PE_LEVEL was found not to have a statistically significant effect on task time for the HT instructions (F (1, 57) = .06, p>.10) for the human translated tasks. This means that when the factor PE_LEVEL is considered without distinctions between languages, there are no statistically significant differences across the two post-editing levels PEz (M=188.38, SE=12.75) and PEp (M=183.53, SE=13.06).

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect on task time (F (2, 57) =.12, p>.10). This means that the factor Language (no distinction between PE_LEVELs) combined with the factor PE_LEVEL (no distinction between languages) do not have a joint effect on total task time.

Table 5:12 shows the mean and standard deviation for each language and their respective post-editing levels (in seconds), while Figure 5:7 illustrates the estimated marginal means for each post-editing level.

Groups		Mean	Std. Deviation
DE	PEz	194.75	35.57
	PEp	181.81	78.46
ZH	PEz	199.33	64.33
	PEp	207.51	48.56
JP	PEz	171.06	78.38
	PEp	161.27	88.22

Table 5:12- Mean and Standard Deviation for Task Time (secs) - HT instructions – Translated Content

A pairwise comparison indicated that there was a moderate statistical difference between the languages JP (M=166.16, SE =13.39) and ZH (M=203.41, SE =15.44) at the p<.10 level. Tukey post-hoc results show that there were no statistically significant differences between JP_PEz and ZH_PEz groups; or JP_PEp and ZH_PEp.

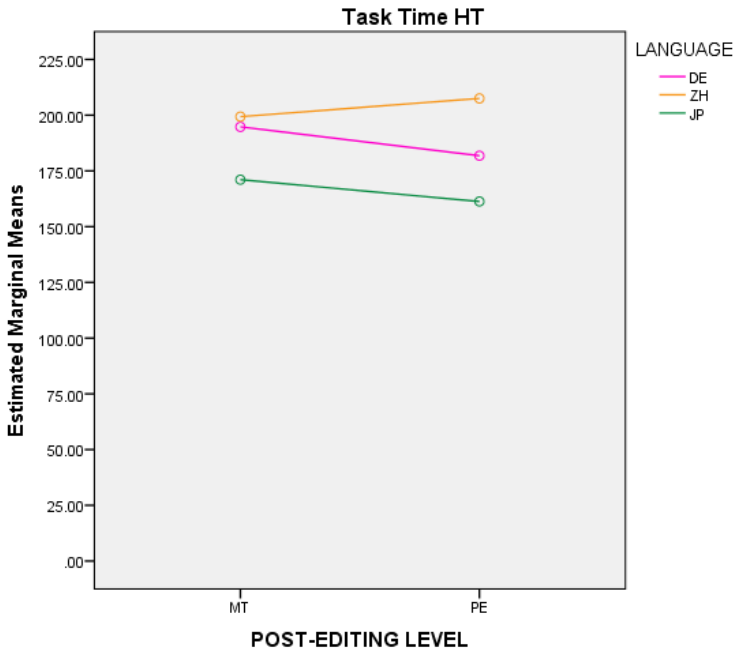


Figure 5:7 - Task Time (secs) - HT Instructions – Translated Content

Overall, there were no statistically significant differences among any of the groups, which indicate that participants from both PEz and PEp groups from all translated languages were comparatively fast at performing the tasks with HT instructions.

Comparison with Source

The performance of the participants who used the English source of the HT Instructions was also computed for task time. Table 5:13 shows the mean and standard deviation for each language and their respective post-editing levels (in seconds) compared to the English source, while Figure 5:8 illustrates the estimated marginal means for each post-editing level compared to the Source.

Groups		Mean	Std. Deviation
EN	SOURCE	130.38	37.59
DE	PEz	194.75	35.57
	PEp	181.81	78.46
ZH	PEz	199.33	64.33
	PEp	207.51	48.56
JP	PEz	171.06	78.38
	PEp	161.27	88.22

Table 5:13 – Mean and Standard Deviation for Task Time (secs) - (HT Instructions - Source

Regarding languages (no distinctions between PE-LEVELs), participants who used the English instructions were statistically faster when compared to the ones who used the ZH instructions (M=203.41, SE=14.82) at the $p<.05$ level and moderately faster when compared to the participants who used the DE (M= 188.28, SE=18.32) instructions at the $p<.10$ level. Regarding PE_LEVEL, there was a statistically significant difference for the English source when compared to the post-editing level PEz (M=188.38, SE=12.24) where $p<.10$.

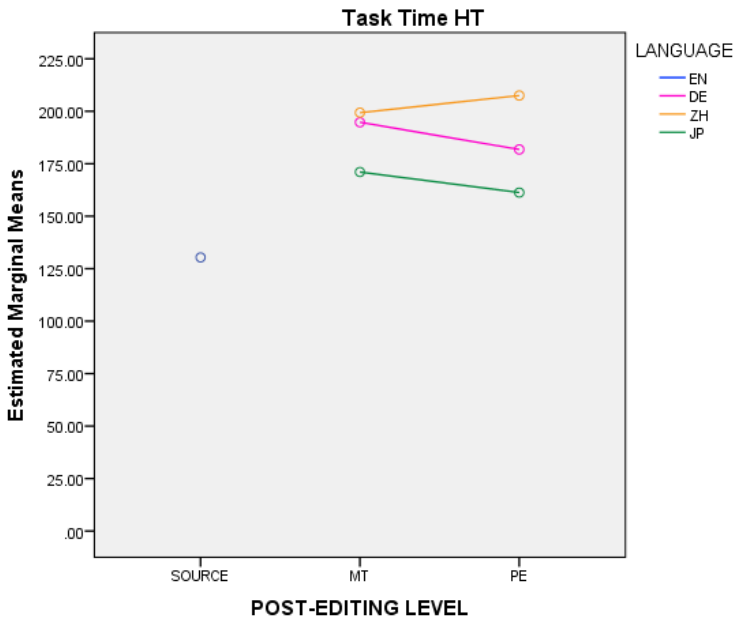


Figure 5:8 - Task Time HT Instructions – Source (secs)

Tukey post-hoc results show that there was no statistical difference between the EN_SOURCE and the other groups DE_PEz, DE_PEp, ZH_PEz, ZH_PEp, JP_PEz and JP_PEp. This indicates that the difference in task time for the participants who used the source of the HT instructions and participants who saw the HT instructions embedded in the PEp and PEz instructions (for DE, ZH and JP) was not statistically significant.

5.1.4.2 Efficiency (Successful Tasks/Task Time)

Efficiency is measured as the number of successful tasks completed (out of all possible tasks) when task time is taken into account. High scores for the efficiency variable indicate greater efficiency.

5.1.4.2.1 *MT Instructions*

A two-way ANOVA was conducted to compare the main effects of the factors Language and PE_LEVEL on total task time for the MT instructions.

LANGUAGE: The factor Language did not have a statistically significant effect on efficiency, where F (2, 57) =.451, p>.10. This means that when the factor Language is considered without distinctions between PE_LEVELs), there is no statistically significant difference across the three translated languages, DE (M= 6.21, SE=0.67), ZH (M= 5.39, SE=0.54), and JP (M= 5.75, SE=0.47) for efficiency.

PE_LEVEL: The factor PE_LEVEL was found to have a very statistically significant effect on efficiency, where F (1, 57) = 17.79, p<.001. This indicates that when the factor PE_LEVEL is considered without distinctions between languages, there is a significant difference across the two post-editing levels PEz (M=4.40, SE=0.45) and PEp (M=7.17, SE=0.48) at the p<.001 level.

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect on efficiency, where F (2, 57) =1.59, p>.10. This means that the factor Language (no distinctions between PE_LEVELs) combined with the factor PE-LEVEL (no distinctions between languages) do not have a joint effect on efficiency.

Table 5:14 shows the mean and standard deviation for each language and their respective post-editing levels, while Figure 5:9 illustrates the estimated marginal means for each post-editing level.

Groups		Mean	Std. Deviation
DE	PEz	4.42	2.53
	PEp	8.00	2.74
ZH	PEz	3.65	1.62
	PEp	7.14	3.26
JP	PEz	5.13	2.26
	PEp	6.39	2.53

Table 5:14 - Mean and Standard Deviation for Efficiency – Translated Content

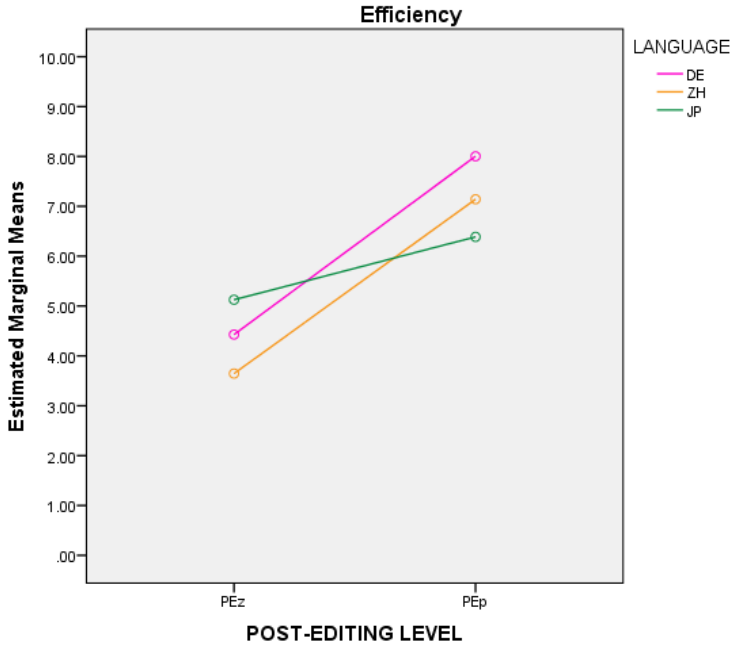


Figure 5:9 - Efficiency – Translated Content

When looking at efficiency across PE-LEVELs, DE_PEP group shows higher efficiency, followed by the ZH-PEp and JP_PEP respectively. However, there were no statistically significant differences among the PEp groups. When looking at the PEz groups, JP_PEz shows higher efficiency followed by the DE_PEz and ZH_PEz groups respectively, however, no statistically significant differences were found among the PEz groups. This indicates that participants from all the PEp groups (DE_PEp, ZH_PEp and JP_PEp) were similarly fast when performing the tasks; as well as participants from all the PEz groups (DE_PEz, ZH_PEz, and JP_PEz).

Regarding PE_LEVEL within languages, the Simplified Chinese language presented a very statistically significant difference between ZH_PEp and ZH_PEz at the $p<.005$ level. A statistically significant difference was also observed for the German language DE_PEp and DE_PEz at the $p<.05$ level. These results indicate that for the Simplified Chinese and German languages, participants who used the lightly post-edited instructions (PEp) were significantly more efficient than participants who used the raw machine translated instructions (PEz). For the Japanese language, however, no statistically significant difference was found between JP_PEp and JP_PEz, where $p>.10$. This indicates that participants who used the PEp instructions were as efficient as the participants who used the PEz instructions for the Japanese language.

Overall, participants who used the PEp instructions to perform the tasks were more efficient than participants who used the PEz instructions. However, this difference is not statistically significant different for the Japanese language.

Comparison with Source

The performance of the participants who used the English source of the MT instructions was also computed. Table 5:15 shows the mean and standard deviation for each language and their respective post-editing levels compared to the English source, while Figure 5:10 illustrates the estimated marginal means for each post-editing level compared to the English source.

A pairwise comparison found statistically significant differences between languages (no distinctions between PE_LEVELs), where English was statistically different from German (M=6.21, SE=0.66) and Japanese (M=5.75, SE=0.46) at the $p<.05$ levels, and from Simplified Chinese (M=5.39, SE=0.53) at the $p<.005$. This indicates that participants who used the English instructions were, in general, more efficient when compared to participants who used the German, Simplified Chinese and Japanese instructions.

Regarding PE_LEVELs (no distinctions between languages), a pairwise comparison found very statistically significant differences between the EN_SOURCE and the PE_LEVEL PEz (M= 4.40, SE=0.44) at the $p<.001$ level. A Tukey post-hoc found that the EN_SOURCE group was statistically different from the PEz groups

DE_PEz (p<.05), JP_PEz (p<.05) and very statistically different from the ZH_PEz group (p<.001). This indicates that participants who used the English source instructions were significantly more efficient than participants who used the PEz instructions of all translated languages.

	Groups	Mean	Std. Deviation
EN	SOURCE	8.43	1.94
DE	PEz	4.42	2.53
	PEp	8.00	2.74
ZH	PEz	3.65	1.62
	PEp	7.14	3.26
JP	PEz	5.13	2.26
	PEp	6.39	2.53

Table 5:15 – Mean and Standard Deviation for Efficiency - Source

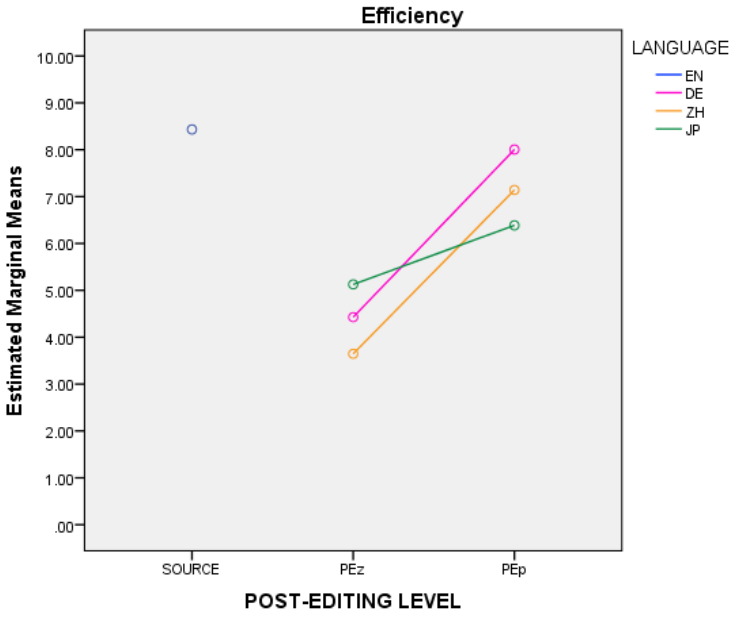


Figure 5:10 - Efficiency - Source

No significant differences were found between the EN_SOURCE group and the PEp groups DE_PEp, ZH_PEp and JP_PEp (p>.10), which indicates that participants who used the PEp instructions of all translated languages were as efficient as participants who used the English source instructions.

Overall, these results indicate that the translated languages groups who used the raw machine translated instructions (PEz) were less efficient when compared to

the participants who used the source instructions. However, the translated languages groups who used the lightly post-edited version of the instructions were as efficient as the group that used the source instructions.

5.1.4.2.2 *HT Instructions*

A two-way ANOVA was conducted to compare the main effects of Language and PE_LEVEL on efficiency for the human translated tasks. Note that this data only involves HT instructions – results for the MT instruction were presented in the previous section (5.1.2.2.1).

LANGUAGE: The factor Language was found to have a statistically significant effect on efficiency, where F (2, 57) = 5.72, p<.005. This indicates when the factor Language is considered without distinctions between PE_LEVELs, there is a statistical difference across the translated content when performing the tasks with the HT instructions – DE (M= 55.01, SE=7.26), ZH (M=45.02, SE =5.88) and JP (M=70.97, SE =5.09).

PE_LEVEL: The factor PE_LEVEL was found not to have a statistically significant effect on efficiency (F (1, 57) = .94, p>.10) for the human translated tasks. This means that when the factor PE_LEVEL is considered without distinctions between languages, there is no statistical differences across the two post-editing levels PEz (M=53.55, SE =4.85) and PEp (M=60.45, SE =5.18).

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect on efficiency for the HT instructions tasks (F (2, 57) =.48, p>.10). This means that the factor Language (no distinctions between PE_LEVELs) combined with the factor PE_LEVEL (no distinctions between Languages) do not have a joint effect on efficiency.

Table 5:16 shows the mean and standard deviation for each language and their respective post-editing levels and Figure 5:11 illustrates the estimated marginal means for each post-editing level.

Groups		Mean	Std. Deviation
DE	PEz	46.96	17.44
	PEp	63.07	24.25
ZH	PEz	46.11	16.84
	PEp	43.94	16.74
JP	PEz	67.60	38.32
	PEp	74.35	30.76

Table 5:16 - Mean and Standard Deviation for Efficiency - HT Instructions – Translated Content

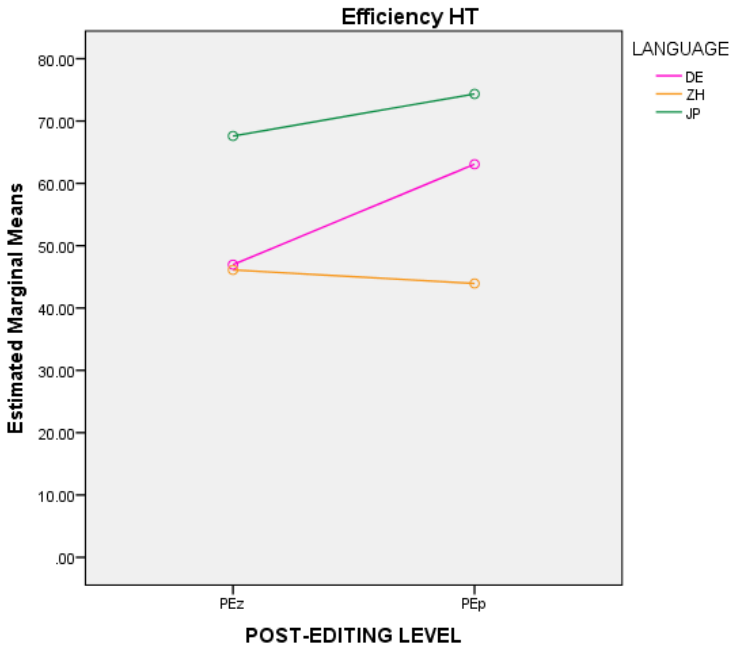


Figure 5:11 - Efficiency HT Instructions – Translated Content

A pairwise comparison showed that there was a very statistically significant difference between the languages (no distinction between PE_LEVELs) JP and ZH at the $p<.005$ level; and also a moderate difference between JP and DE at the $p<.10$ level.

When looking at PE_LEVEL PEp, JP_PEp shows higher efficiency when compared to DE_PEp and ZH_PEp. A statistically significant difference was found for the comparison JP_PEp against the ZH_PEp groups, at the $p<.05$ level, which indicates that Japanese participants who used the PEp instructions where significantly more efficient at performing the HT tasks than the Chinese participants who used the PEp instructions. Regarding the PE_LEVEL PEz, the JP_PEz group also

shows higher efficiency when compared to the DE_PEz and ZH_PEz groups. These results were statistically significant for both comparisons at the $p<.10$ level.

When analysing PE_LEVELs within the three languages, there were no statistically significant differences between the PE_LEVELs for Japanese (JP_PEz and JP_PEp), Chinese (ZH_PEz and ZH_PEp) and for German (DE_PEz and DE_PEp), where $p>.10$.

Overall, these results indicate that the Japanese participants from both groups (JP_PEz and JP_PEp) were in general more efficient at performing tasks using the HT instructions than the other languages groups. However, the lack of difference between PE_LEVELs within which language indicates that participants from both PEz and PEp groups were comparatively efficient at performing the tasks with the HT instructions.

Comparison with Source

The performance of the participants who used the English source of the HT instructions was also computed. Table 5:17 shows the mean and standard deviation for each language and their respective post-editing levels compared to the English source, while Figure 5:12 illustrates the estimated marginal means for each post-editing level compared to the source.

Groups		Mean	Std. Deviation
EN	SOURCE	81.03	17.51
DE	PEz	46.96	17.44
	PEp	63.07	24.25
ZH	PEz	46.11	16.84
	PEp	43.94	16.74
JP	PEz	67.60	38.32
	PEp	74.35	30.76

Table 5:17 - Mean and Standard Deviation for Efficiency HT Instructions - Source

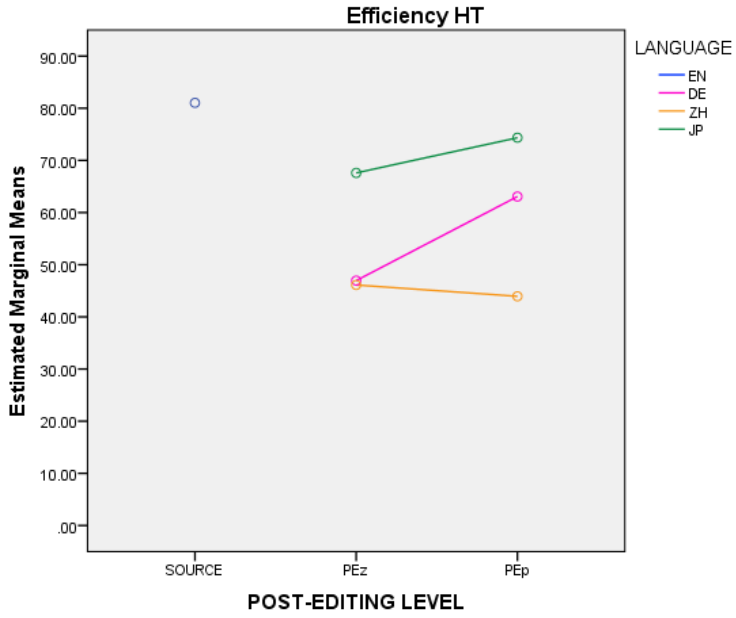


Figure 5:12- Efficiency HT Instructions - SOURCE

The participants who used the English instructions were statistically more efficient at performing tasks 4 and 8 when compared to the participants who used the Simplified Chinese instructions (p<.005) and the participants who used the German instructions (p<.05), without making a distinction between their PE_LEVELs. There was no statistically significant difference between the English and Japanese languages (p>.10).

Regarding PE_LEVELS, a pairwise comparison found statistically significant differences between the English source and PE_LEVEL PEz (M=53.55, SE =4.69) and PEp (M=60.45, SE =5.01) at the p<.05 level. Regarding the PE-LEVEL PEz, the EN_SOURCE group was moderately different from the ZH_PEz group (p<.10), which indicates that participants who used the English source instructions were more efficient than participants who used the ZH_PEz instructions. No differences were found between the English source and the JP_PEz and DE_PEz groups (p>.10).

When analysing the PE-LEVEL PEp, a statistically significant difference was found between the EN_SOURCE group and the ZH_PEp group (p<.05), which indicates that participants who used the English source were strongly more efficient than the participants from the ZH_PEp group. No significant differences were found between the EN_SOURCE and the DE_PEp and JP_PEp groups (p>.10), which indicates these two groups were as efficient the EN_SOURCE group.

Overall, these results indicate that English participants using the source text, where statistically more efficient than German and Chinese participants who used the human translated version of the source.

5.1.5 Cognitive Data

Cognitive data was gathered during the performance of the usability tasks and, therefore, it involved the same participants as the usability tasks. However, because cognitive data analysis requires the eye-tracking recordings to be of good quality, all recordings that presented less than 80% estimated quality were discarded (See section 4.2.1.1.). In consequence, fewer suitable recordings remained to be used for cognitive analysis, making the sample size for the cognitive data analysis smaller: Thirteen for the German language (seven for PEz, six for PEp); fifteen for the Simplified Chinese language (seven for PEz, eight for PEp); and fourteen for the Japanese language (seven for PEz, seven for PEp).

For the analysis of the cognitive data, two areas of interest (AOI) are considered: the AOI instruction (INST) and the AOI user interface (UI). The AOI INST, refers to the window that displayed the instructions (PEz, PEp and SOURCE) the participants read in order to perform the tasks. The AOI UI refers to the window that displayed the user interface of the spreadsheet application.

As repeated measures design is used to assess the cognitive measures (fixation duration, fixation count, visit duration and visit count), the within-subject factor combines the measures for both AOIs. For example, the fixation duration measure is divided into: fixation duration gathered in the instructions AOI (FD_INST) and fixation duration gathered in user interface AOI (FD_UI). When the within-subject factor is defined for fixation duration, both FD_INST and FD_UI are grouped, creating the FD factor. Therefore, when reporting the within-subject factor, the abbreviations are used: FD, FC (fixation count), VD (visit duration), and VC (visit count). When reporting fixation and visits for each AOI, the full name (‘fixation duration’, ‘visit count’) is used.

5.1.5.1 Fixation Duration

As described in Chapter 4, Section 4.4.1.1, fixation duration is the sum of the fixation lengths for all participants divided by the number of all fixations within and AOI.

5.1.5.1.1 *Baseline*

As described in Chapter 4, Section 4.2.2.1, a short text was selected as the baseline for the cognitive data.

As seen in Figure 5:13 the German PEz group has longer fixations on the AOI baseline when compared to the other groups. However, no statistically significant differences were found for any of the groups compared. This lack of significantly differences indicates that in general, all groups had the same level of cognitive effort required when reading the text.

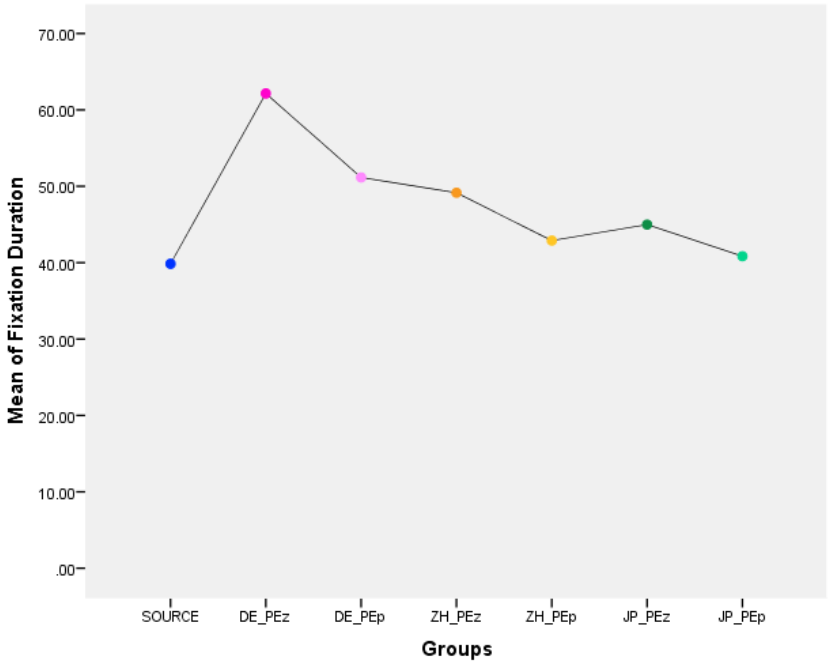


Figure 5:13 - Fixation Duration Baseline - All Groups

5.1.5.1.2 *MT Instructions*

A two-way MANOVA with repeated measures was conducted in order to compare whether the factors Language and PE_LEVEL have an effect on Fixation Duration (FD) for both AOIs: Instruction (FD_INST) and User Interface (FD_UI).

LANGUAGE: The factor Language was found not to have a statistically significant difference on FD, where (F (2, 35) = .19, p>.10). This means that when the factor language is considered without distinctions between PE_LEVELs, there is no statistically significant differences across the three translated languages DE (M=502.70, SE=56.98), ZH (M=544.61, SE=53.00), JP (M=504.06, SE=56.98).

POST-EDITING LEVEL: The factor PE_LEVEL was also found not to have a statistically significant difference on FD, where (F (1, 35) = .33, p>.10). This means that when the factor PE_LEVEL is considered without distinctions between languages, there is no statistically significant differences across the two post-editing levels PEz (M=535.84, SE=44.70) and PEp (M=498.40, SE=46.22).

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect on fixation duration, where (F (2, 35) = .75, p>.10). This means that the factor language combined with the factor PE_LEVEL do not have a joint effect on FD.

Table 5:18 shows the mean and standard deviation for each language and their respective post-editing levels (in seconds) for each AOI (instructions and UI).

AOIs	Groups		Mean	Std. Deviation
FD_INST	DE	PEz	441.46	46.81
		PEp	440.60	95.24
	ZH	PEz	484.08	198.97
		PEp	349.17	104.26
	JP	PEz	403.36	191.21
		PEp	346.81	157.04
FD_UI	DE	PEz	591.23	147.08
		PEp	537.54	228.60
	ZH	PEz	742.59	355.87
		PEp	602.63	303.77
	JP	PEz	552.37	245.51
		PEp	713.70	356.91

Table 5:18 - Mean and Standard Deviation for Fixation Duration (secs) - Translated Content

The test of within-subjects determined that there was a statistically significant difference between FD_INST and FD_UI (F (1, 35) =55 .93, p<.001), where shorter fixations can be observed in the AOI INST (M=410.91, SE=22.43) when compared to the AOI UI (M=623.34, SE=44.35). Figure 5:14 illustrates the estimated marginal

means for each post-editing level for fixation duration instructions, while Figure 5:15 illustrates for fixation duration UI.

When looking at the AOI INST across PE_LEVELs (Figure 5:14), all PEp groups present shorter fixation duration when compared to their PEz groups, which indicates that the groups which used the post-edited instruction had less cognitive effort observed when reading the instructions. However, these results are only statistically significant for the Simplified Chinese language (ZH_PEz INST (M= 484.08, SD=198.97), ZH_PEp INST (M=349.17, SD=104.26)).

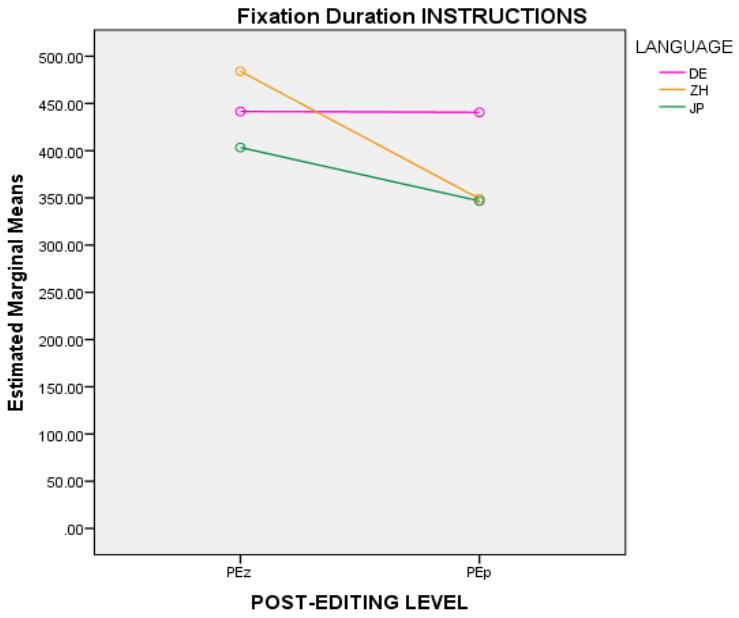


Figure 5:14 - Fixation Duration Instructions (secs) – Translated Content

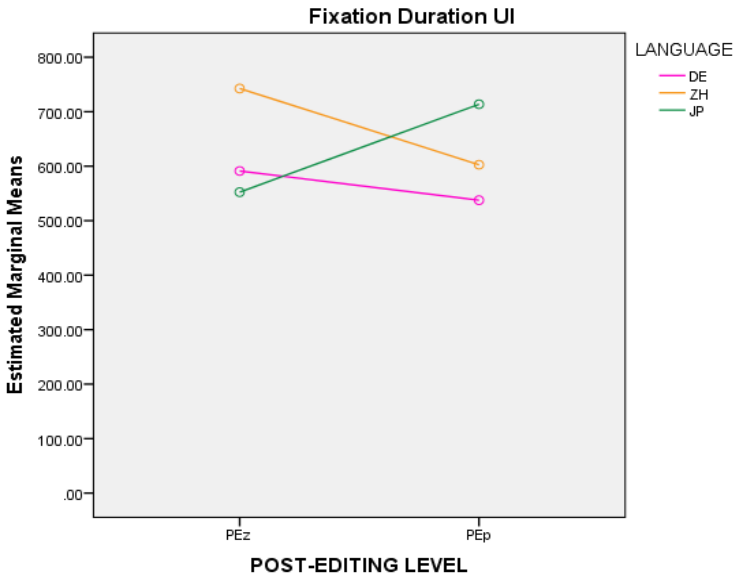


Figure 5:15 - Fixation Duration UI (secs) – Translated Content

When looking at the AOI UI across PE_LEVELS (Figure 5:15), the JP_PEp group has longer fixation duration when compared to JP_PEz, which indicates that the participants of the JP_PEp group spend more time in fixations in the user interface window performing the task than in the instructions window reading the instructions and, therefore, more cognitive effort can be observed in the AOI UI. This result differs from the Chinese and German languages where the longer fixation duration for the UI is seen in the PEz groups. However, none of these results (for Japanese, Chinese and German) were found to be statistically significant.

Figure 5:16 shows the differences between FD_INST and FD_UI for each group. When comparing both AOIs (INST *vs* UI), all groups have longer fixation duration in the AOI user interface when compared to the AOI instruction. These results were statistically significant for the groups DE_PEz, ZH_PEz, ZH_PEp, JP_PEz, and JP_PEp at the $p<.05$ level; apart from the DE_PEp group which was not statistically significant at the $p>.10$ level. This means that for all groups, there was more cognitive effort related to the user interface window against the instruction window. This is an expected result since the UI is larger and contains more details to be looked at.

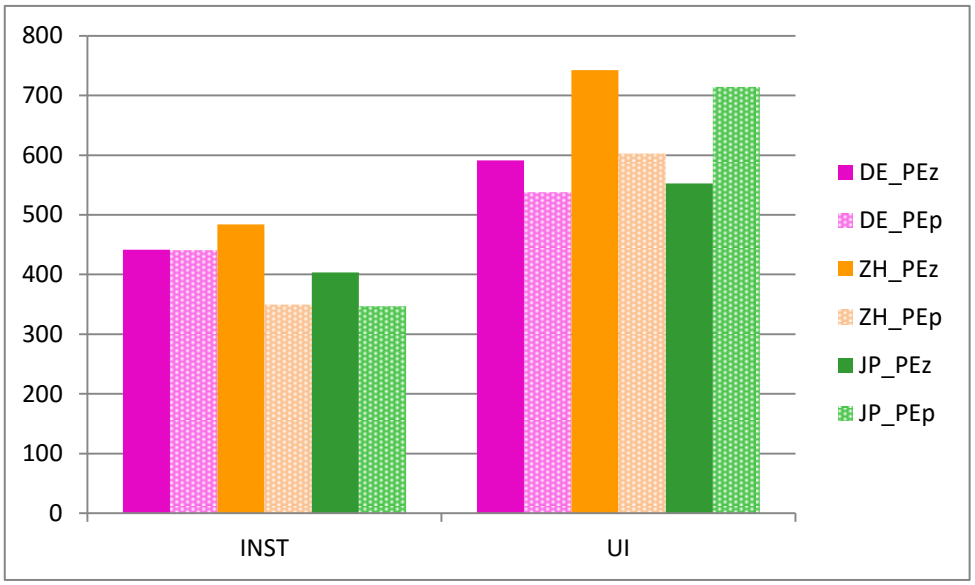


Figure 5:16 – Differences per group for Fixation Duration (secs) – Translated Content

Comparison with Source

The performance of the participants who used the English source of the MT Instructions was also computed for fixation duration via a one-way MANOVA with repeated measures. Table 5:19 shows the mean and standard deviation for each language and their respective post-editing levels (in seconds) for each AOI (instructions and UI) compared to the English source.

AOIs	Groups		Mean	Std. Deviation
FD_INST	EN	SOURCE	278.57	45.73
	DE	PEz	441.46	46.81
		PEp	440.60	95.24
	ZH	PEz	484.08	198.97
		PEp	349.17	104.26
	JP	PEz	403.36	191.21
		PEp	346.81	157.04
FD_UI	EN	SOURCE	400.65	96.84
	DE	PEz	591.23	147.08
		PEp	537.54	228.6
	ZH	PEz	742.59	355.87
		PEp	602.63	303.77
	JP	PEz	552.37	245.51
		PEp	713.7	356.91

Table 5:19 - Mean and Standard Deviation for Fixation Duration (secs) – Source

The factor PE_LEVEL[31] was found not to have a statistically significant effect on fixation duration (F (6, 42) = 1.27, p>.10). The test of within-subjects determined that FD did not have any significant effects in the interaction with PE_LEVEL (F (6, 42) = 1.85, p>.10). However, there was a statistically significant difference between FD_INST and FD_UI (F (1, 42) =55 .93, p<.001).

A pairwise comparison found that the participants who used the source instructions (EN (M=339.61, SE=71.62)) had shorter FD when compared to the DE_PEz group (M=516.34, SE=71.62), at the p <.10 level; ZH_PEz (M=613.33, SE=71.62), at the p <.05 level; and JP_PEp (M=512.28, SE=71.62), at the p<.10 level.

Figure 5:17 illustrates the estimated marginal means for each language and their PE_LEVEL compared to the Source for the instructions AOI, while Figure 5:18

[31] When comparing cognitive data against the source, PE_LEVEL here means all groups including the source. It facilitates to compute the one-way MANOVA and reporting the plots. It is important to remember that the group 'source' uses the English version of instructions and does not undergo any translation or post-editing. Therefore, PE_LEVEL here refers to PEz, PEp and Source.

shows the means for each language and their PE_LEVEL compared to the source for the AOI user interface.

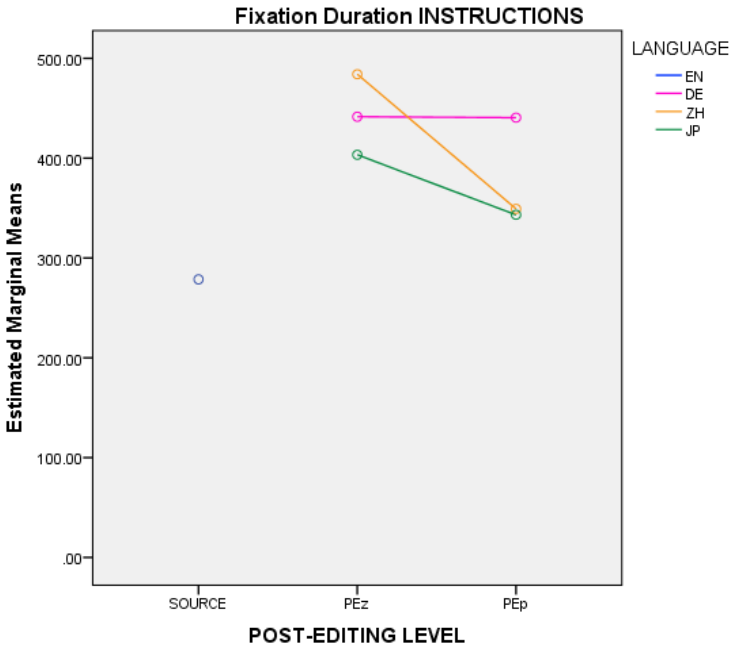


Figure 5:17 - Fixation Duration Instructions (secs) - Source

When looking at the AOI INST across groups (Figure 5:17), the EN_Source group presents the shortest FD in the instruction when compared to all the groups. However, these effect was statistically significant only against the DE_PEz, DE_PEp (p<.05), ZH_PEz (p<.005), and JP_PEz (p<.10) groups.

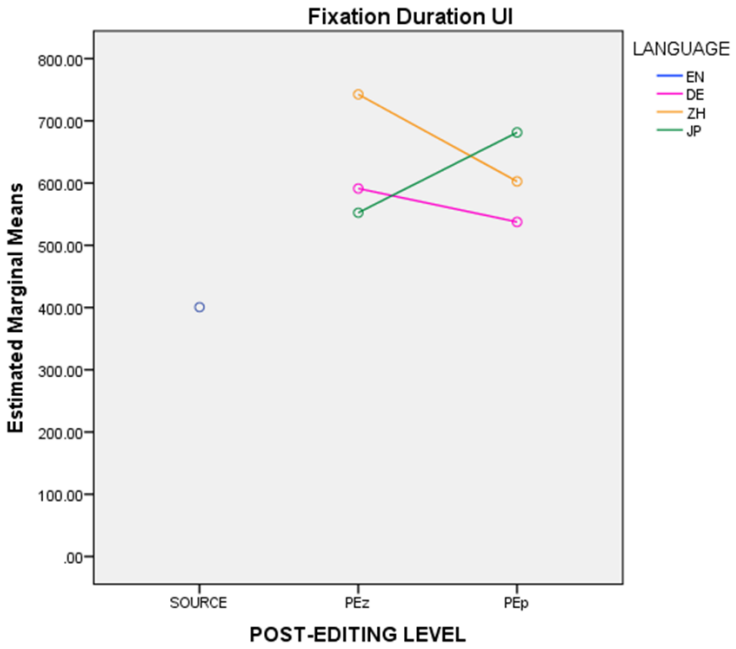


Figure 5:18 - Fixation Duration UI (secs) - Source

When looking at the AOI UI across groups (Figure 5:18), the EN_Source group presents shorter FD in the UI when compared to all the groups. However, this effect was statistically significant only against the ZH_PEz and JP_PEp groups at the p<.005 level.

Figure 5:19 shows the differences between FD_INST and FD_UI for each group compared to the EN_Source. The source also follows the previous results in which all the groups present higher fixation duration in the AOI UI when compared to the AOI INST. This result was statistically significant for the EN_Source group at the p <.05 level. This means that for all groups, including the EN_Source group, there was more cognitive effort related to the user interface window against the instruction window.

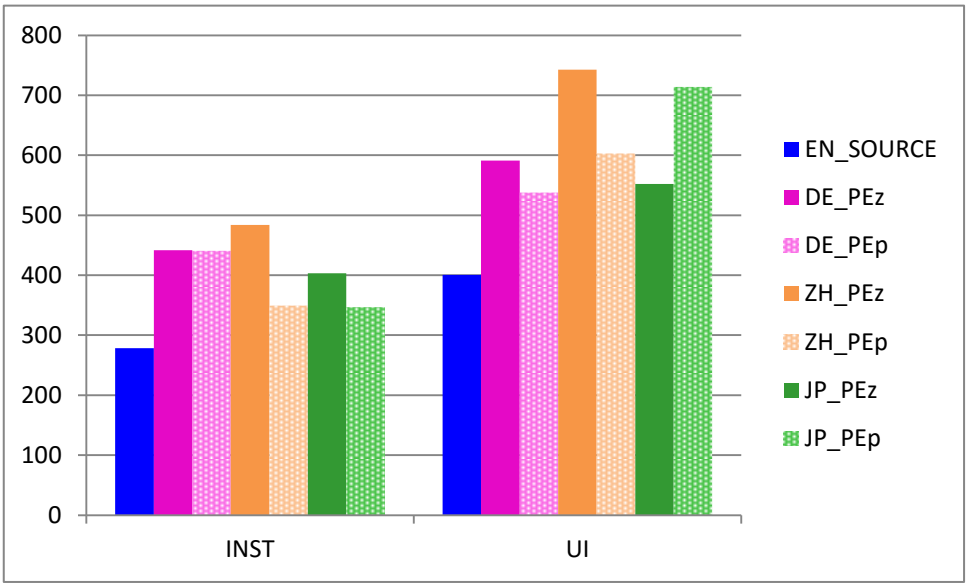


Figure 5:19 - Differences per group for Fixation Duration (secs) – Source

5.1.5.1.3 *HT Instructions*

A two-way MANOVA with repeated measures was conducted in order to compare whether the factors Language and PE_LEVEL have an effect on Fixation Duration (FD) for both AOIs: Instruction (FD_INST) and User Interface (FD_UI) for the HT Instructions.

LANGUAGE: The factor Language was found not to have a statistically significant difference on FD, where (F (2, 35) = 1.97, p>.10). This means that when the factor language is considered without distinctions between PE_LEVELs, there is no statistically significant differences across the three translated languages DE (M=88.12, SE=9.83), ZH (M=97.49, SE=9.14), JP (M=71, SE=9.83).

POST-EDITING LEVEL: The factor PE_LEVEL[32] was also found not to have a statistically significant difference on FD, where (F (1, 35) = .17, p>.10). This means that when the factor PE_LEVEL is considered without distinctions between languages, there is no statistically significant differences across the two post-editing levels PEz (M=87.87, SE=7.71), and PEp (M=83.20, SE=7.97).

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect on FD, where (F (2, 35) = .04, p>.10). This means that

[32] PE_LEVEL here refers to the groups (PEp and PEz) which had the HT instructions embedded.

the factor language combined with the factor PE_LEVEL do not have a joint effect on FD.

Table 5:20 shows the mean and standard deviation for each language and their respective post-editing levels per AOI (instructions and UI) for the HT instructions.

AOIs	Groups		Mean	Std. Deviation
FD_INST	DE	PEz	86.51	11.54
		PEp	94.88	27.67
	ZH	PEz	95.68	30.41
		PEp	76.86	13.80
	JP	PEz	68.46	33.72
		PEp	62.17	30.55
FD_UI	DE	PEz	96.91	26.40
		PEp	74.20	48.84
	ZH	PEz	106.20	53.12
		PEp	111.25	32.28
	JP	PEz	73.50	49.28
		PEp	79.89	72.55

Table 5:20 - Mean and Standard Deviation for Fixation Duration (secs) – HT Instructions - Translated Content

The test of within-subjects determined that there was a statistically significant difference between FD_INST and FD_UI (F (1, 35) = 3.87, p<.10), where shorter fixations can be observed in the AOI INST (M=80.76, SE=4.02) when compared to the AOI UI (M=90.32, SE=7.56). Figure 5:20 illustrates the estimated marginal means for each post-editing level for fixation duration instructions, while Figure 5:21 illustrates for fixation duration UI for the HT Instructions.

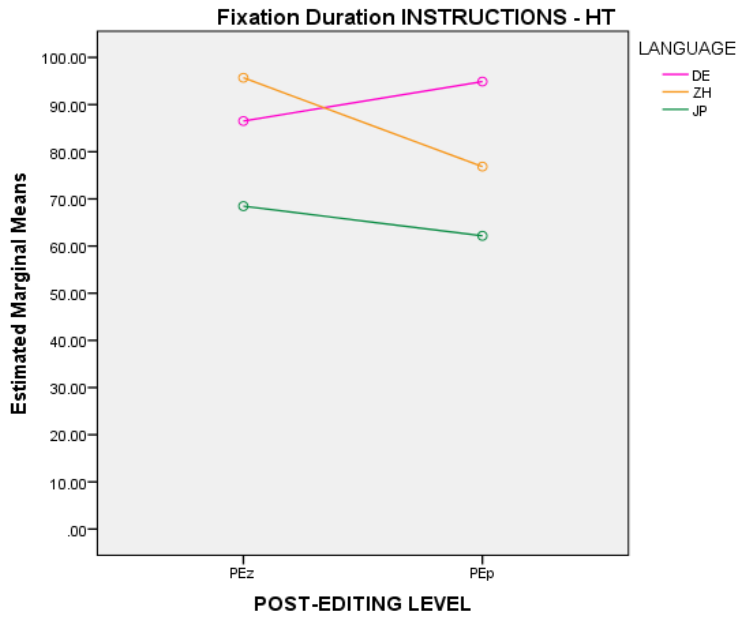


Figure 5:20 - Fixation Duration Instructions (secs) - HT Instructions - Translated Content

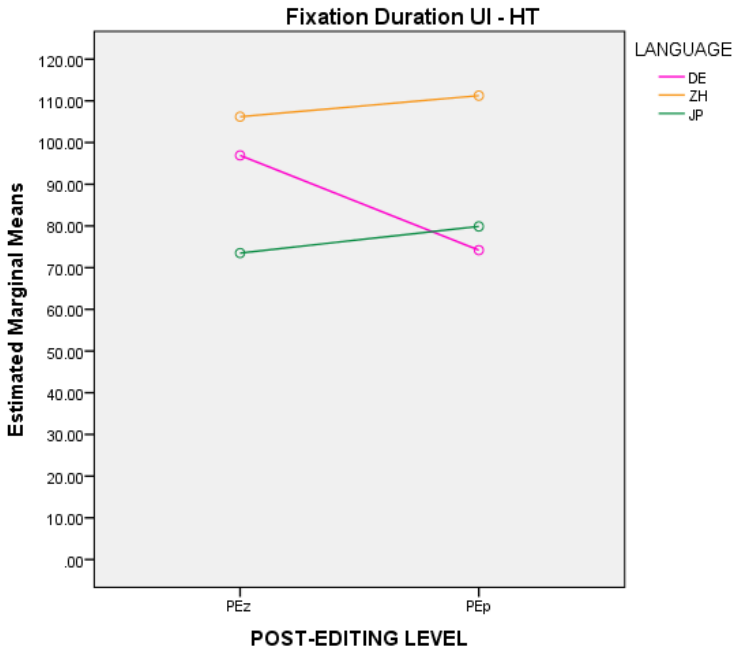


Figure 5:21 - Fixation Duration UI (secs) – HT Instructions - Translated Content

When looking at the AOI INST across PE_LEVELs (Figure 5:20), longer fixations can be observed for the ZH_PEz and JP_PEz groups when compared to their PEp groups. This is interesting because it indicates that the groups which used the PEz instructions (to perform the MT tasks) had more cognitive effort observed when reading the human translated instructions. The German language interestingly

shows longer fixations in the AOI INST for the PEp group when compared to the PEz group when using the HT instructions. However, neither results were statistically significant (p >.10).

When looking at the AOI UI across PE_LEVELS (Figure 5:21), longer fixations can be observed for the ZH_PEp and JP_PEp groups when compared to their PEz groups, which indicates that group who used the HT instructions embedded in the PEp instructions had more cognitive effort observed when using the UI window. The German language, again, in contrast to the other languages, shows longer fixations for the PEz group when using HT instructions. However, neither results were statistically significant (p >.10).

Figure 5:22 shows the differences between FD_INST and FD_UI for each language and post-editing level for the AOIs instructions and UI for the HT instructions. Apart from the DE_PEp groups, all the other groups have longer fixation time in the AOI user interface when compared to the AOI instruction. These results, however, are statistically significant only for the ZH_PEp group at the p<.005 level. For the DE_PEp group, which present longer fixations time in the AOI Instruction, no statistically significant difference was found when comparing to the UI AOI at the p >.10 level.

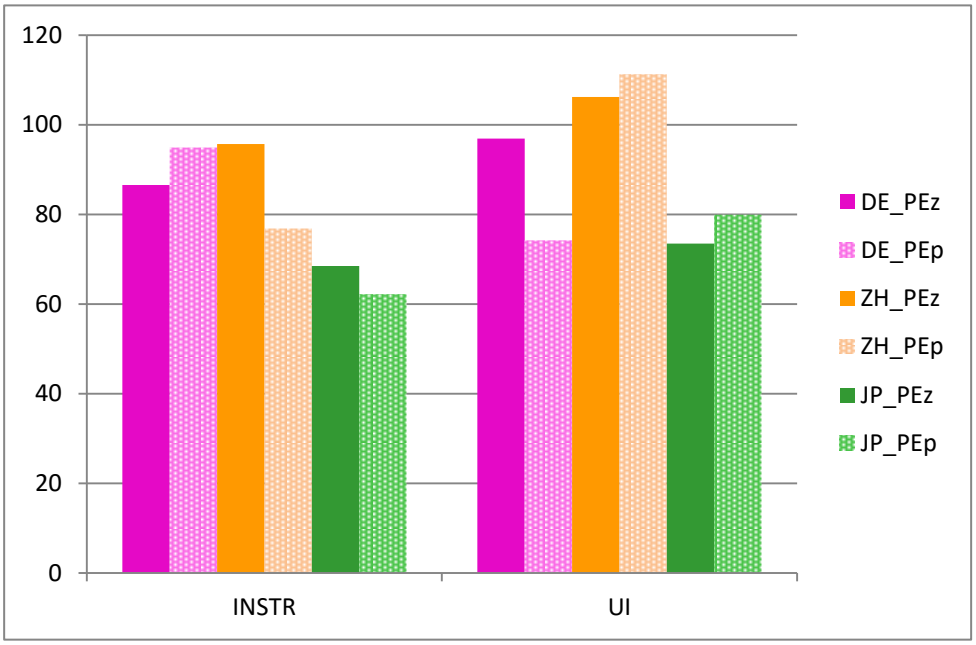


Figure 5:22 – Differences per group for Fixation Duration (secs) – HT Instructions – Translated Content

Comparison with Source

The performance of the participants who used the English source of the HT Instructions was also computed for fixation duration via a one-way MANOVA with repeated measures.

Table 5:**21** shows the mean and standard deviation for each language and their respective post-editing levels (in seconds) for each AOI (instructions and UI) compared to the English source.

AOIs	Groups		Mean	Std. Deviation
FD_INST	EN	SOURCE	53.44	19.76
	DE	PEz	86.51	11.54
		PEp	94.88	27.67
	ZH	PEz	95.68	30.41
		PEp	76.86	13.80
	JP	PEz	68.46	33.72
		PEp	62.17	30.55
FD_UI	EN	SOURCE	60.14	24.92
	DE	PEz	96.91	26.40
		PEp	74.20	48.84
	ZH	PEz	106.20	53.12
		PEp	111.25	32.28
	JP	PEz	73.50	49.28
		PEp	79.89	72.55

Table 5:21 - Mean and Standard Deviation for Fixation Duration (secs) - Source – HT instructions

The factor PE_LEVEL[33] was found not to have a statistically significant effect on FD (F (6, 42) = 1.59, p>.10). The test of within-subjects determined that FD (when both AOIs are considered) had a significant effect in the interaction with PE_LEVEL (F (6, 42) = 2.19, p<.10). There was also a statistically significant difference between FD_INST and FD_UI (F (1, 42) = 4.69, p<.05). Figure 5:23 illustrates the estimated marginal means for each language and their PE_LEVEL compared to the Source for the AOI instructions, while Figure 5:24 shows the means for each language and their PE_LEVEL compared to the source for the AOI user interface.

[33] Again, it is important to stress that when comparing cognitive data against the source, PE_LEVEL means all groups including the source, as it facilitates to compute the one-way MANOVA and reporting the plots.

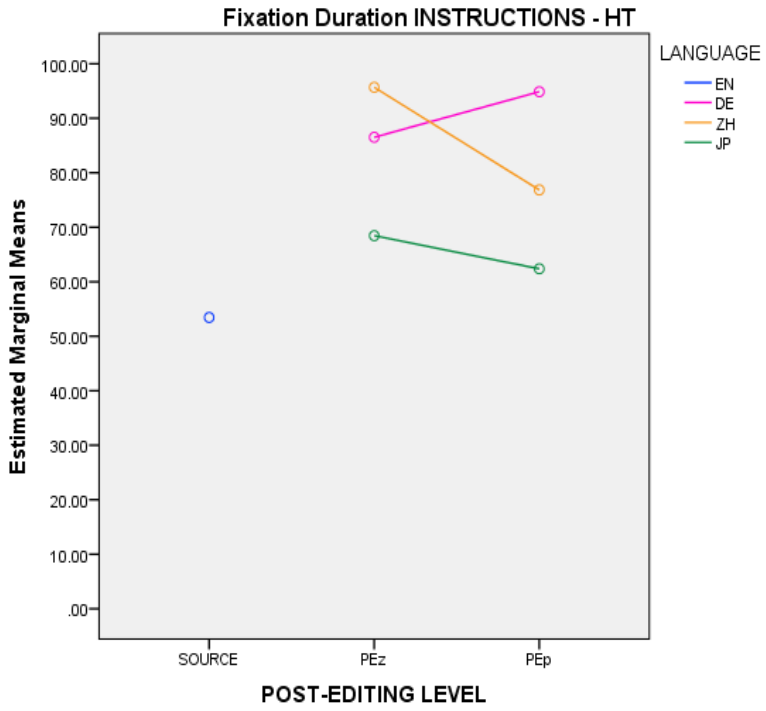


Figure 5:23 - Fixation Duration Instructions (secs) – HT Instructions – Source

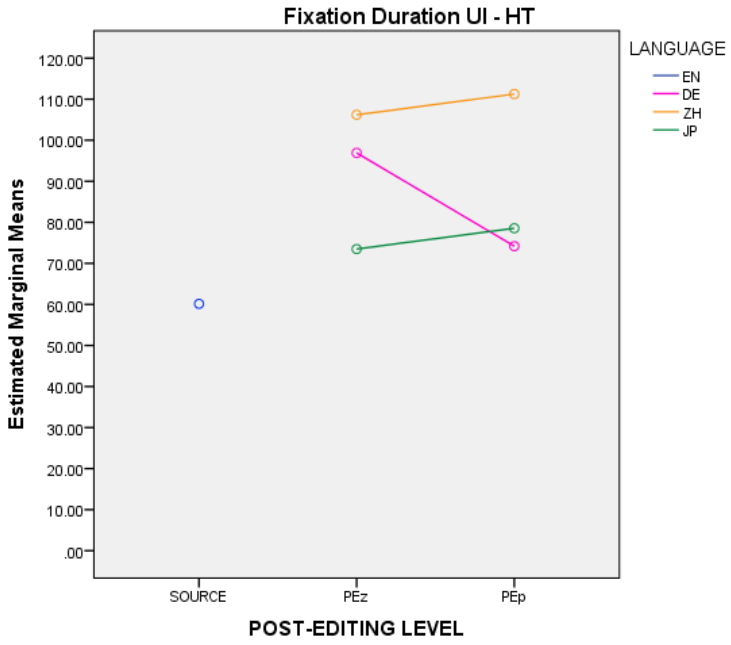


Figure 5:24 - Fixation Duration UI (secs) – HT Instructions – Source

A pairwise comparison found that the participants who used the source instructions (EN (M=56.79, SE=12.57)) had shorter FD (both AOIS) when compared to the DE_PEz group (M=91.71, SE=12.57), at the *p* <.10 level; ZH_PEz (M=100.94, SE=12.57), at the *p* <.05 level; and ZH_PEp (M=94.05, SE=11.76), at the *p*<.05 level.

When looking at the AOI INST across groups (Figure 5:23), the EN_Source group presents shorter fixations in the instruction when compared to all the groups. However, this effect was statistically significant only against the DE_PEz (p<.05), DE_PEp (p<.005), ZH_PEz (p<.005), and ZH_PEp (p<.10) groups.

When looking at the AOI UI across groups (Figure 5:24), the EN_Source group presents shorter fixation time in the UI when compared to all the groups. However, this effect was statistically significant only against the ZH_PEz (p<.10) and ZH_PEp groups at the p<.05 level.

Figure 5:25 shows the differences between FD_INST and FD_UI for each group compared to the EN_Source. The source also follows the previous results in which all the groups (apart from DE_PEp) present higher fixation duration in the UI AOI when compared to the INST UI. This result, however, was not statistically significant for the EN_Source group at the p >.10 level.

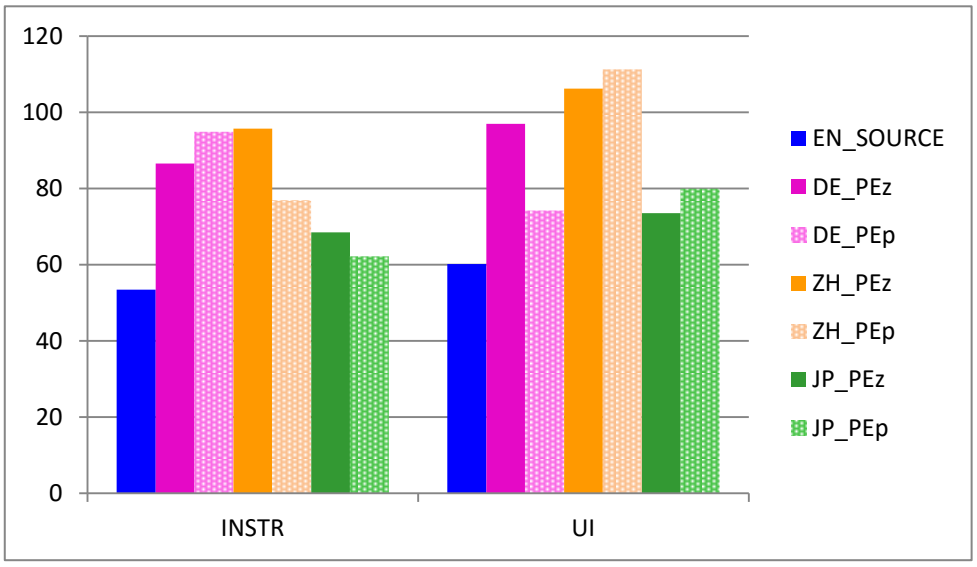


Figure 5:25 - Differences per group for FD – Source – HT Instructions

5.1.5.2 Fixation Count

As described in Chapter 4, Section 4.4.1.1, fixation count (FC) is the total number of fixations within an AOI.

5.1.5.2.1 *Baseline*

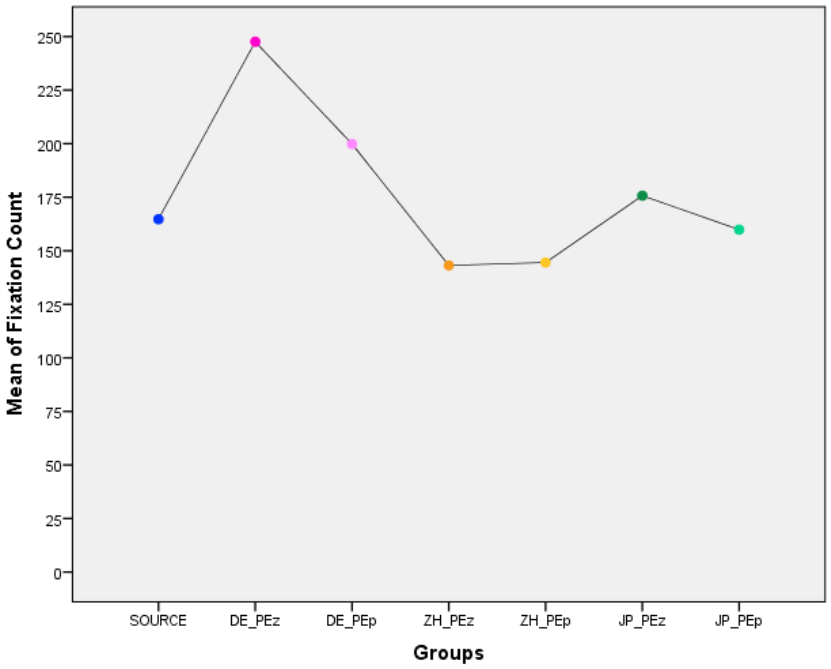


Figure 5:26 - Fixation Count Baseline - all groups

When looking at number of fixations in the AOI baseline (Figure 5:26), it is possible to notice that the German PEz group has more fixations on the AOI baseline when compared to the other groups. This result was statistically significant when compared to the ZH groups and the JP_PEz groups. However, no statistically significant differences were found for any when comparing PEz and PEp groups within languages, as well as when comparing EN_Source against the PEp and PEz levels. This lack of significantly differences between PEp, PEz and source, indicates that in general, all groups had the same level of cognitive effort required when reading the text.

5.1.5.2.2 *MT Instructions*

A two-way MANOVA with repeated measures was conducted in order to compare whether the factors Language and PE_LEVEL have an effect on Fixation Count (FC) for both AOIs: Instruction (FC_INST) and User Interface (FC_UI).

LANGUAGE: The factor Language was found not to have a statistically significant difference on FC, where (F (2, 35) = .14, p>.10). This means that when the factor language is considered without distinctions between PE_LEVELs, there is no

statistically significant differences across the three translated languages DE (M=1628.45, SE=145.79), ZH (M=1523.68, SE=135.62), JP (M=1562.77, SE=145.79).

POST-EDITING LEVEL: The factor PE_LEVEL was also found not to have a statistically significant difference on FC, where (F (1, 35) = .61, p>.10). This means that when the factor PE_LEVEL is considered without distinctions between languages, there is no statistically significant differences across the two post-editing levels PEz (M=1635.90, SE=114.37), and PEp (M=1507.36, SE=118.27).

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect on FC, where (F (2, 35) = .38, p>.10). This means that the factor language combined with the factor PE_LEVEL do not have a joint effect on FC.

Table 5:22 shows the mean and standard deviation for each language and their respective post-editing levels for each AOI (instructions and UI).

AOIs	Groups		Mean	Std. Deviation
FC_INST	DE	PEz	1526.57	179.21
		PEp	1545.50	368.54
	ZH	PEz	1283.43	350.33
		PEp	1044.13	275.20
	JP	PEz	1419.00	573.32
		PEp	1249.00	487.45
FC_UI	DE	PEz	1761.57	319.84
		PEp	1680.17	739.91
	ZH	PEz	2091.29	740.03
		PEp	1675.88	874.02
	JP	PEz	1733.57	670.66
		PEp	1849.50	748.67

Table 5:22 - Mean and Standard Deviation for Fixation Count - Translated Content

The test of within-subjects determined that there was a statistically significant difference between FC_INST and FC_UI (F (1, 35) = 41.17, p<.001), where fewer fixations can be observed on the AOI INST (M=1344.60, SE=61.07) when compared to the AOI UI (M=1798.66, SE=110.94). Figure 5:27 illustrates the estimated marginal means for each post-editing level for fixation count instructions, while Figure 5:28 illustrates for fixation count UI.

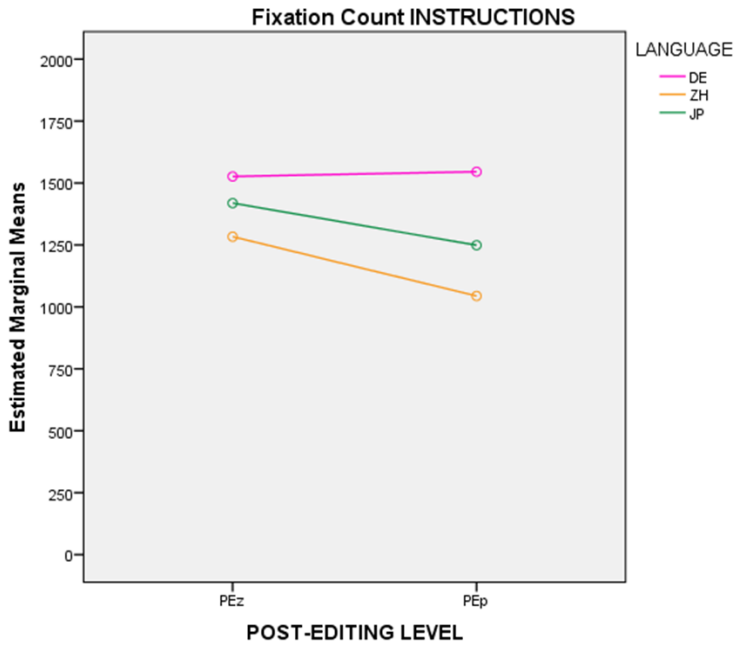


Figure 5:27 - Fixation Count Instructions - Translated Content

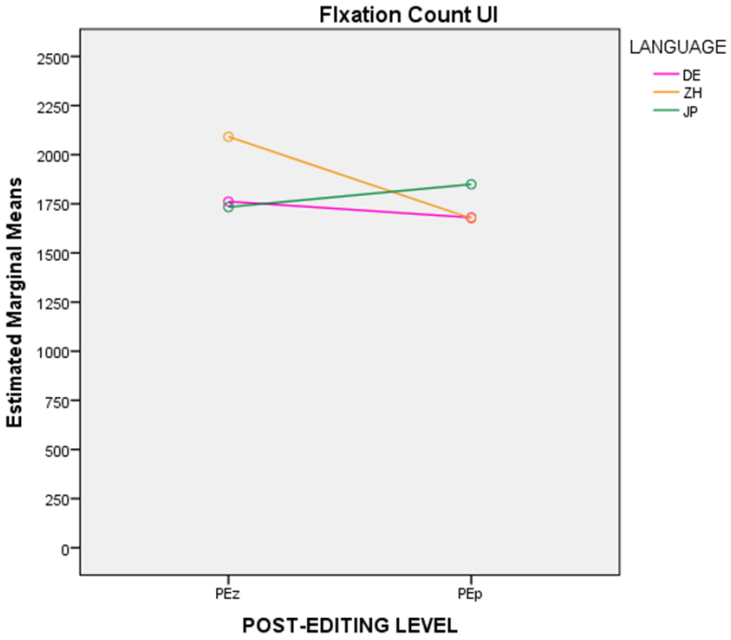


Figure 5:28 - Fixation Count UI - Translated Content

When looking at the AOI INST across PE_LEVELs (Figure 5:27), both ZH_PEp and JP_PEp groups present fewer fixation count in the AOI INST when compared to their PEz groups. DE_PEp has slightly more fixation counts in the AOI INST when

compared to its PEz group. However, none of the differences for all groups were statistically significant.

When looking at the AOI UI across PE_LEVELs, again, the JP_PEp group has greater fixation count when compared to JP_PEz, which differs from the German and Simplified Chinsese languages where fixation count in the UI is higher for the PEz groups. However, these results were found not to be statistically significant.

Figure 5:29 shows the differences between FC_INST and FC_UI for each language and post-editing level. When comparing both AOIS (INST *vs* UI) across groups, all groups presented more fixation count in the AOI user interface when compared to the AOI instructions, which is expected since it is in the UI that participants perform the task and it also contains more details than the AOI instructions. These results were statistically significant for ZH_PEz, JP_PEz, ZH_PEp and JP_PEp groups at the $p<.05$ level. Neither German groups (DE_PEz, DE_PEp) had statistically significant differences on fixation count between the AOIs instructions and UI, at the $p>.10$ level.

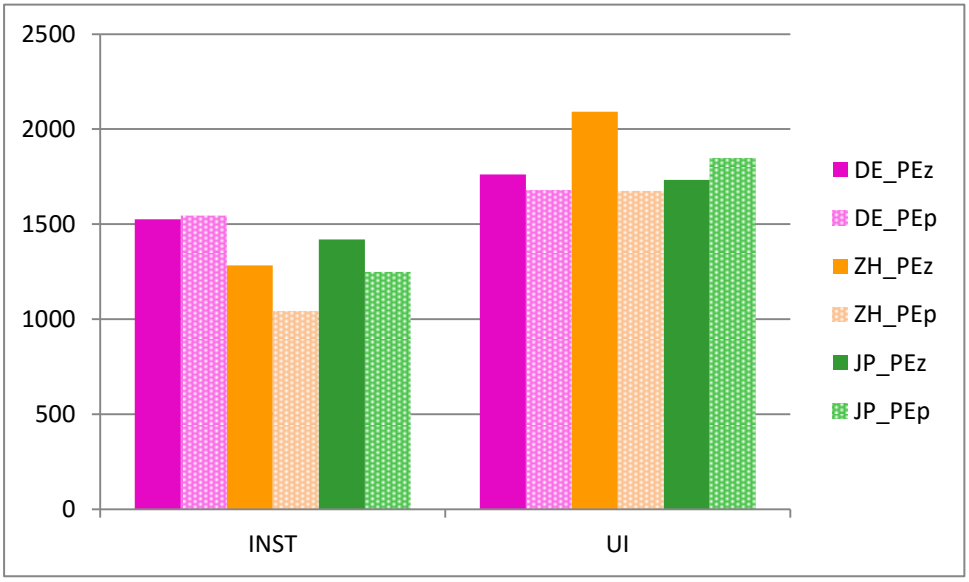


Figure 5:29 - Differences per PE_Level and Language for Fixation Count – Translated Content

Comparison with Source

The performance of the participants who used the English source of the MT Instructions was also computed for fixation count via a one-way MANOVA with

repeated measures. Table 5:23 shows the mean and standard deviation for each language and their respective post-editing levels (in seconds) for each AOI (instructions and UI) compared to the English source.

AOIs	Groups		Mean	Std. Deviation
FD_INST	EN	SOURCE	1017.14	153.29
	DE	PEz	1526.57	179.21
		PEp	1545.5	368.54
	ZH	PEz	1283.43	350.33
		PEp	1044.13	275.2
	JP	PEz	1419,00	573.32
		PEp	1249,00	487.45
FD_UI	EN	SOURCE	1297.00	302.02
	DE	PEz	1761.57	319.84
		PEp	1680.17	739.91
	ZH	PEz	2091.29	740.03
		PEp	1675.88	874.02
	JP	PEz	1733.57	670.66
		PEp	1849.50	748.67

Table 5:23 - Mean and Standard Deviation for Fixation Count – Source

The factor PE_LEVEL was found not to have a statistically significant effect on FC (F (6, 42) = 1.06, p>.10). The test of within-subjects determined that FC has a significant effect in the interaction with PE_LEVEL (F (6, 42) = 2.26, p>.10). There was also a statistically significant difference between FC_INST and FC_UI (F (1, 42) = 47.77, p<.001).

A pairwise comparison found that the participants who used the source instructions (EN (M=1157.07, SE=183.41)) had fewer FC (both AOIS) when compared to the DE_PEz (M=1644.07, SE=183.41) and DE_PEp (M=1612.83, SE=198.10) groups at the p<.10 level; ZH_PEz (M=1687.35, SE=183.41) at the p <.05 level.

Figure 5:30 illustrates the estimated marginal means for each language and their PE_LEVEL compared to the Source for the AOI instructions, while Figure 5:31 shows the means for each language and their PE_LEVEL compared to the source for the AOI user interface.

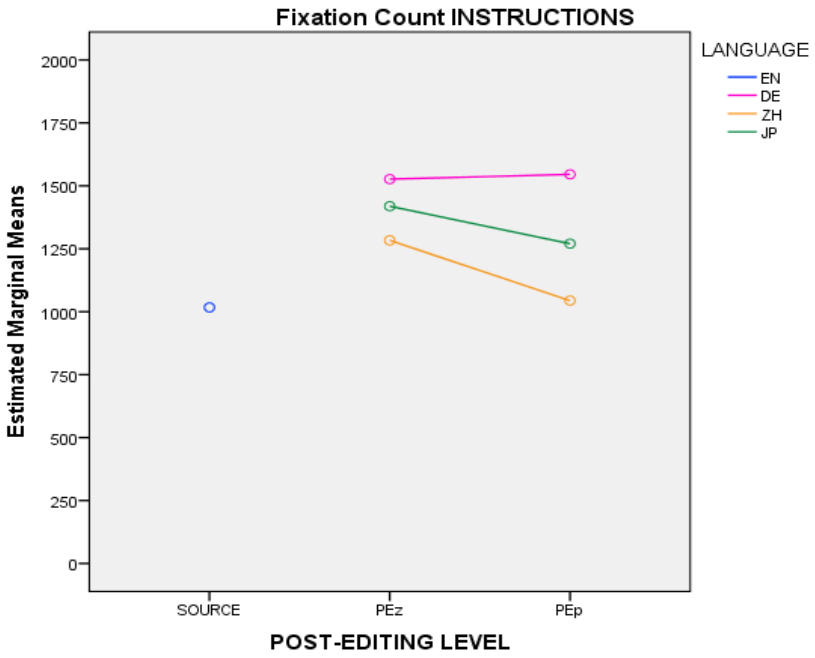


Figure 5:30 - Fixation Count Instructions - Source

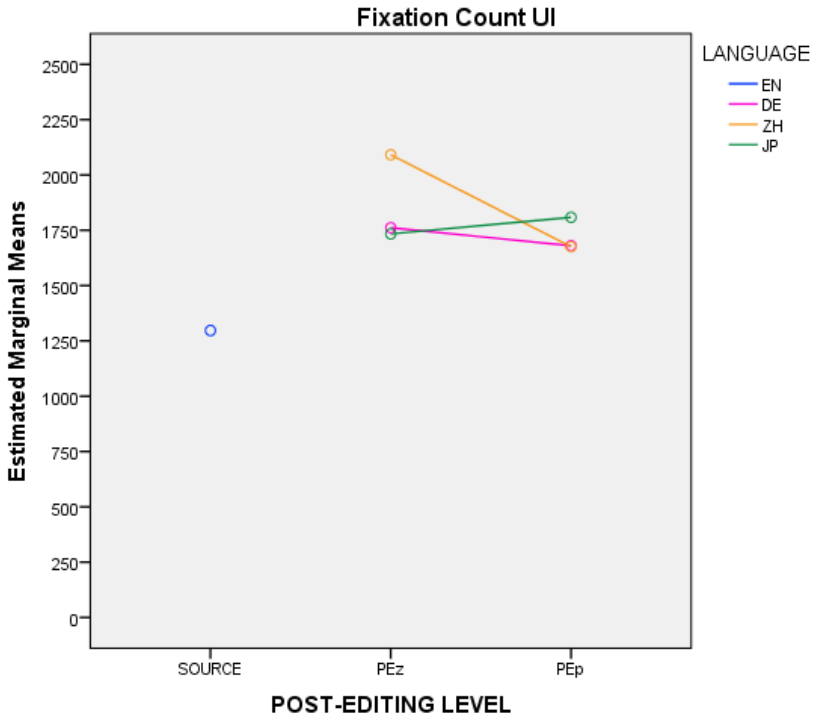


Figure 5:31 - Fixation Count UI - Source

When looking at the AOI INST across groups (Figure 5:30), the EN_Source group presents fewer fixations in the instruction when compared to all the groups. However, these effect was statistically significant only against the DE_PEz, DE_PEp and JP_PEz (p<.05) groups.

When looking at the AOI UI across groups (Figure 5:31), the EN_Source group presents fewer fixations in the UI when compared to all the translated language groups. However, this effect was statistically significant only against the ZH_PEz group at the $p<.05$ level.

Figure 5:32 shows the differences between fixation count in the AOI INST and in the AOI UI for each group compared to the EN_Source. The result for the source is in line with the previous results in which all the translated language groups present greater fixation count in the AOI UI when compared to the AOI INST. This result was statistically significant for the EN_Source group at the p <.10 level. This means that for all groups, including for the EN_Source group, there was more cognitive effort related to the user interface window against the instruction window, which again, is an expected result since the task has to be performed in the UI.

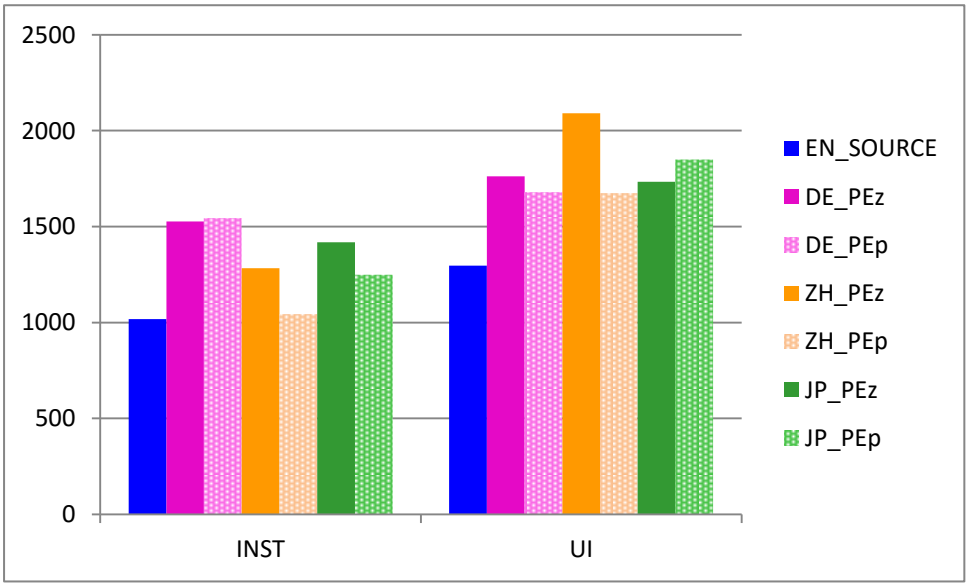


Figure 5:32 - Differences per group for Fixation Count - Source

5.1.5.2.3 *HT Instructions*

A two-way MANOVA with repeated measures was conducted in order to compare whether the factors Language and PE_LEVEL have an effect on Fixation Count (FC) for both AOIs: Instruction (FC_INST) and User Interface (FC_UI) for the HT Instructions.

LANGUAGE: The factor Language was found not to have a statistically significant difference on FC, where (F (2, 35) = .68, p>.10). This means that when the factor language is considered without distinctions between PE_LEVELs, there is no statistically significant differences across the three translated languages DE (M=279.38, SE=27.41), ZH (M=254.83, SE=25.50), JP (M=234.11, SE=27.41).

POST-EDITING LEVEL: The factor PE_LEVEL was also found not to have a statistically significant difference for FC, where (F (1, 35) = .15, p>.10). This means that when the factor PE_LEVEL is considered without distinctions between languages, there is no statistically significant differences across the two post-editing levels PEz (M=262.09, SE=21.50), and PEp (M=250.12, SE=22.24).

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect on FC, where (F (2, 35) = .03, p>.10). This means that the factor language combined with the factor PE_LEVEL do not have a joint effect on FC.

Table 5:24 shows the mean and standard deviation for each language and their respective post-editing levels per AOI (instructions and UI) for the HT instructions.

AOIs	Groups		Mean	Std. Deviation
FC_INST	DE	PEz	300.00	63.08
		PEp	325.83	110.42
	ZH	PEz	238.71	55.38
		PEp	211.63	27.50
	JP	PEz	249.71	103.95
		PEp	229.00	94.66
FC_UI	DE	PEz	279.71	80.55
		PEp	212.00	141.42
	ZH	PEz	272.86	115.55
		PEp	296.13	70.78
	JP	PEz	231.57	142.49
		PEp	226.17	183.02

Table 5:24 - Mean and Standard Deviation for Fixation Count - HT Instructions - Translated Content

The test of within-subjects determined that there was not a statistically significant difference between FC_INST and FC_UI (F (1, 35) = 3.22, p>.10), which indicates that both AOIs INST (M=259.14, SE=12.39) and UI (M=253.07, SE=19.55) presented the roughly the same number of fixations. Figure 5:33 illustrates the

estimated marginal means for each post-editing level for fixation count instructions, while Figure 5:34 illustrates for fixation count UI for the HT Instructions.

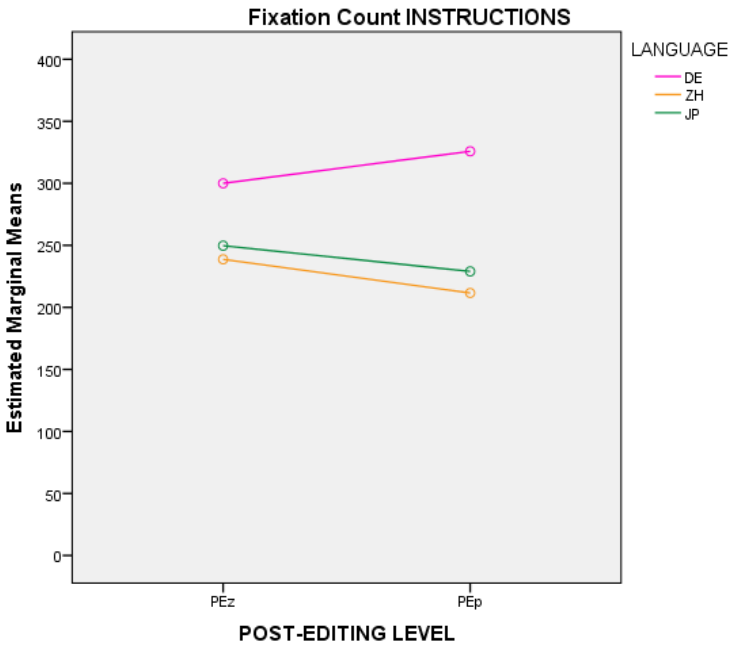


Figure 5:33 - Fixation Count Instructions - HT Instructions - Translated Content

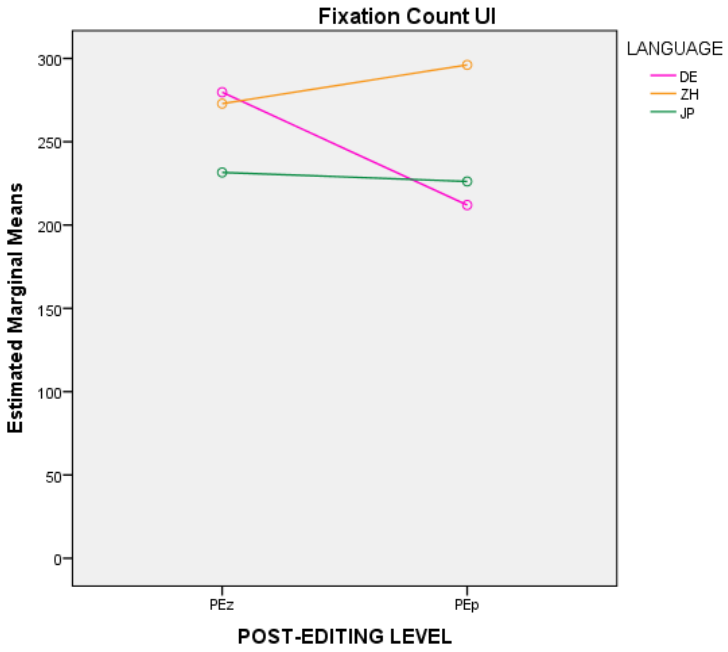


Figure 5:34 - Fixation Count UI - HT Instructions - Translated Content

When looking at the AOI INST across PE_LEVELs (Figure 5:33), a higher number of fixations can be observed for the ZH_PEz and JP_PEz groups when compared to their PEp groups, which indicates that the ZH and JP groups who used

the raw machine translated version of the instructions required more cognitive effort when reading the instructions, when fixation count is taken as a measure of cognitive effort. The German language interestingly shows more fixations in the AOI INST group for the PEp group when compared to the PEz group. However, neither results were statistically significant (p >.10).

When looking at the AOI UI across PE_LEVELS (Figure 5:34), a higher number of fixations can be observed for the DE_PEZ and JP_PEz groups when compared to their PEp groups, which indicates that the groups which used the raw MT version of the instructions had more cognitive effort observed when using the UI window. The Chinese language, oppositely from the other languages, shows more fixations for the PEp group. However, none of the results were statistically significant (p >.10).

Figure 5:35 shows the differences between fixation count for each language and post-editing level for the AOIs instructions and UI. Apart from the Simplified Chinese groups, all the other groups have higher fixation count in the instructions AOI when compared to the user interface AOI. These results, however, are statistically significant only for the DE_PEp group at the p<.005 level. For the ZH_PEp groups, which presents more fixations in the user interface AOI a statistically significant difference was found when comparing to the INST AOI at the p <.005 level.

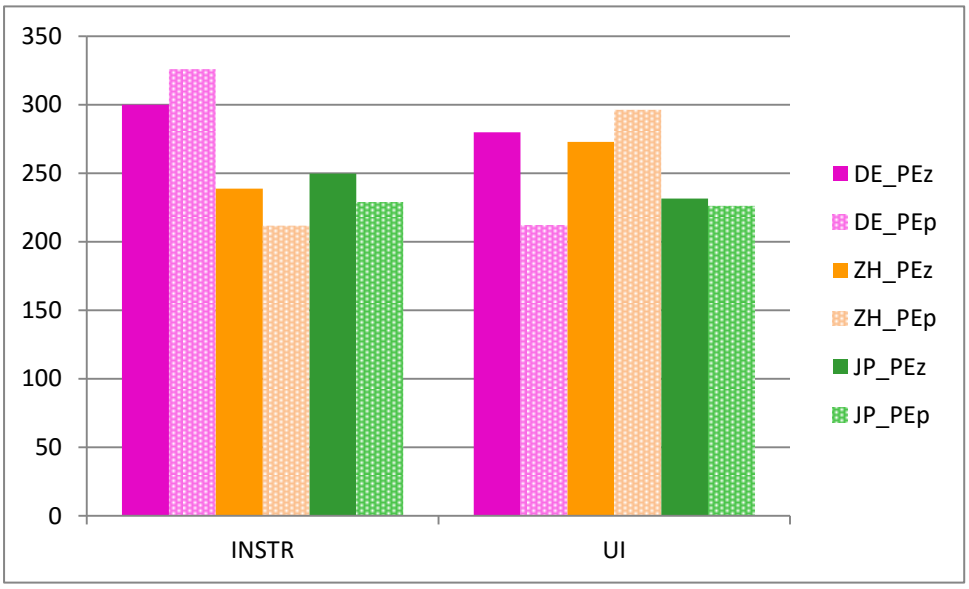


Figure 5:35 - Differences per group for FC - HT Instructions - Translated Content

Comparison with Source

The performance of the participants who used the English source of the HT Instructions was also computed for fixation count via a one-way MANOVA with repeated measures. Table 5:25 shows the mean and standard deviation for each language and their respective post-editing levels (in seconds) for each AOI (instructions and UI) compared to the English source.

The factor PE_LEVEL was found not to have a statistically significant effect on FC (F (6, 42) = .79, p>.10). The test of within-subjects determined that FC (when both AOIs are considered) had a significant effect in the interaction with PE_LEVEL (F (6, 42) = 5.72, p<.005). There was no statistically significant difference between FC_INST and FC_UI (F (1, 42) = .50, p>.10). A pairwise comparison found that the participants who used the source instructions (EN (M=189.78, SE=35.25)) had fewer FC (both AOIS) when compared to the DE_PEz group (M=289.85, SE=35.25), at the p <.05 level.

AOIs	Groups		Mean	Std. Deviation
FD_INST	EN	SOURCE	191.14	59.350
	DE	PEz	300.00	63.08
		PEp	325.83	110.42
	ZH	PEz	238.71	55.38
		PEp	211.63	27.50
	JP	PEz	249.71	103.95
		PEp	229.00	94.66
FD_UI	EN	SOURCE	188.43	78.079
	DE	PEz	279.71	80.55
		PEp	212.00	141.42
	ZH	PEz	272.86	115.55
		PEp	296.13	70.78
	JP	PEz	231.57	142.49
		PEp	226.17	183.02

Table 5:25 - Mean and Standard Deviation Fixation Count - HT Instructions – Source

Figure 5:36 illustrates the estimated marginal means for each language and their PE_LEVEL compared to the Source for the AOI instructions, while Figure 5:37 shows the means for each language and their PE_LEVEL compared to the source for the AOI user interface.

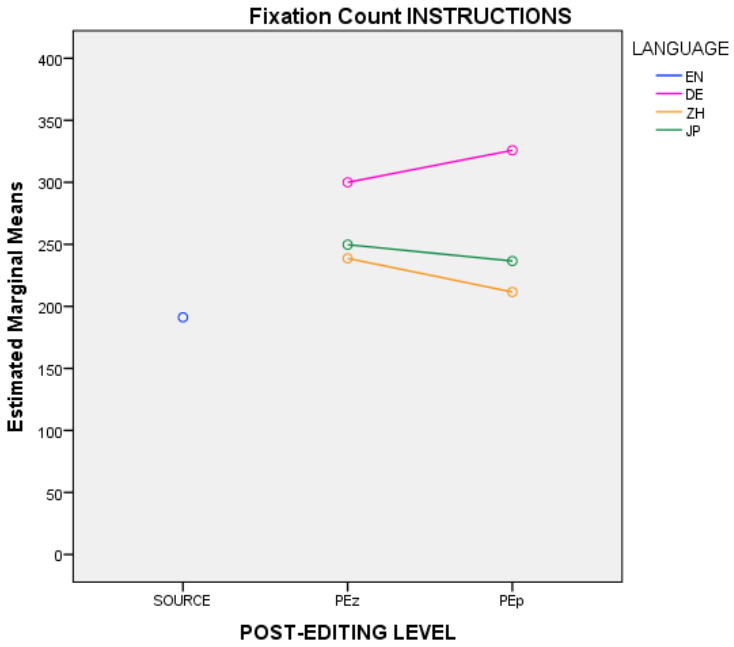


Figure 5:36 - Fixation Count Instructions - HT Instructions - Source

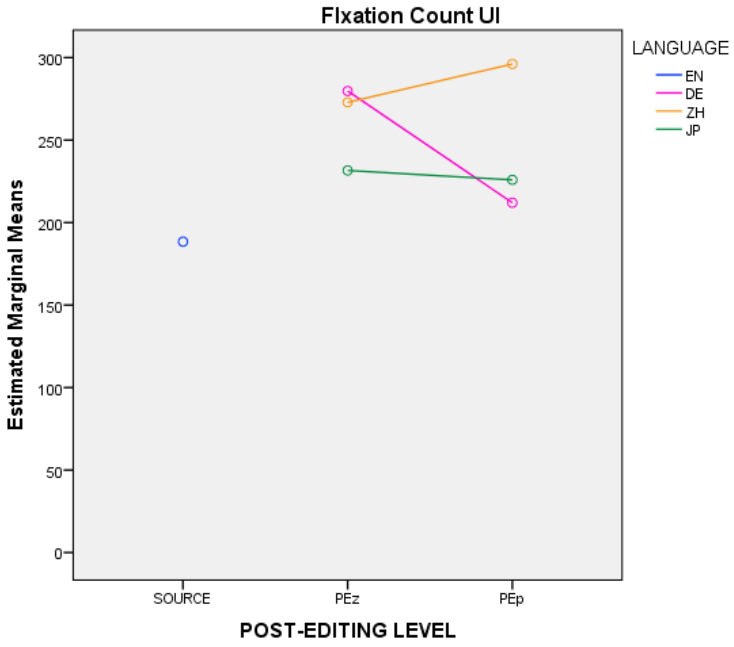


Figure 5:37 - Fixation Count UI - HT Instructions - Source

When looking at the AOI INST across groups (Figure 5:36), the EN_Source group presents fewer fixations in the instruction when compared to all the groups. However, this effect was statistically significant only against the DE_PEz (p<.05) and DE_PEp (p<.005) groups.

When looking at the AOI UI across groups (Figure 5:24), the EN_Source group presents fewer fixations in the UI when compared to all the groups. However, this effect was statistically significant only against the ZH_PEp group at the $p<.10$ level.

Figure 5:38 shows the differences between FC_INST and FC_UI for each group compared to the EN_Source. The source also follows the previous results in which the German and Japanese groups present greater fixation count in the AOI INST when compared to the INST UI (apart from the Simplified Chinese group). This result, however, was not statistically significant for the EN_Source group at the $p>.10$ level.

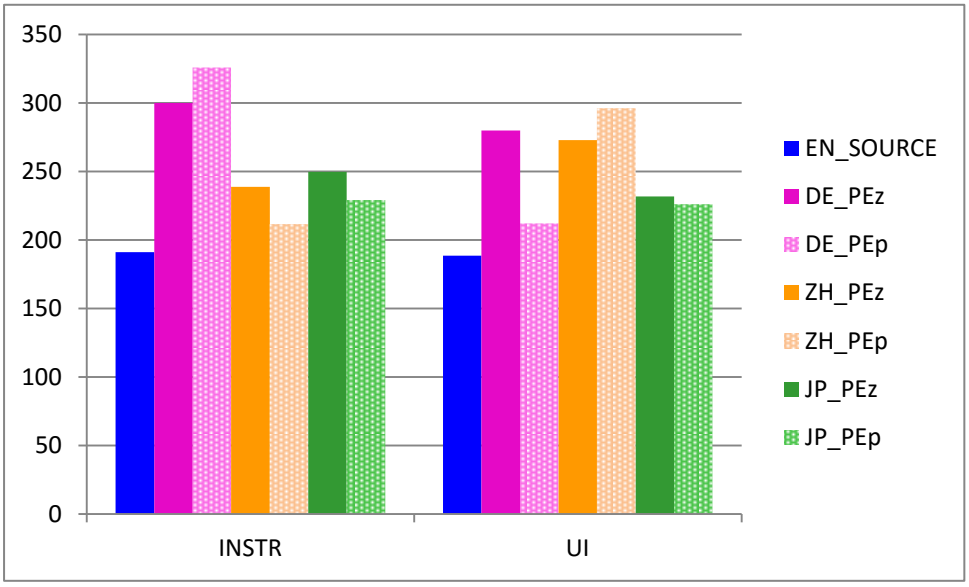


Figure 5:38 - Differences per group for FC - HT Instructions - Source

5.1.5.3 Visit Duration

As described in Chapter 4, section 4.4.1.1, visit duration is the length of a visit within and AOI. The results for the visit duration were similar to the results of fixation duration and, therefore, in order to make this chapter more readable and avoid repetitions, the results will not be reported here. The full report of the results for visit duration can be found in Appendix D.

5.1.5.4 Visit Count

As described in Chapter 4, Section 4.4.1.1, visit count (VC) is the number shifts of attention between AOIs.

5.1.5.4.1 *MT Instructions*

A two-way MANOVA with repeated measures was conducted in order to compare whether the factors Language and PE_LEVEL have an effect on Visit Count (VC) for both AOIs: Instruction (VC_INST) and User Interface (VC_UI).

LANGUAGE: The factor Language was found not to have a statistically significant difference on VC, where (F (2, 35) = .39, p>.10). This means that when the factor language is considered without distinctions between PE_LEVELs, there is no statistically significant differences across the three translated languages DE (M=177.95, SE=13.04), ZH (M=189.75, SE=12.13), JP (M=193.65, SE=13.04).

POST-EDITING LEVEL: The factor PE_LEVEL was also found to have a statistically significant effect on VC, where (F (1, 35) = 4.04, p<.10). This means that when the factor PE_LEVEL is considered without distinctions between languages, there is a statistically significant differences across the two post-editing levels PEz (M=201.90, SE=10.23), and PEp (M=172.33, SE=10.58) for visit count (in both AOIs).

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect on VC, where (F (2, 35) = 2.41, p>.10). This means that the factor language combined with the factor PE_LEVEL do not have a joint effect on VC.

Table 5:26 shows the mean and standard deviation for each language and their respective post-editing levels for each AOI (visit count instructions and visit count UI).

AOIs	Groups		Mean	Std. Deviation
VC_INST	DE	PEz	159.86	33.54
		PEp	155.83	43.89
	ZH	PEz	208.14	53.36
		PEp	148.38	27.89
	JP	PEz	200.57	62.96
		PEp	166.83	46.71
VC_UI	DE	PEz	190.43	60.19
		PEp	205.67	47.39
	ZH	PEz	242.71	61.01
		PEp	159.75	32.02
	JP	PEz	209.71	66.85
		PEp	197.50	67.27

Table 5:26 - Mean and Standard Deviation for Visit Count - Translated Content

The test of within-subjects determined that there was a statistically significant difference between VC_INST and VC_UI (F (1, 35) = 17.13, p<.001), where higher visits can be observed to the AOI UI (M=200.96, SE=8.87) against the AOI INST (M=173.26, SE=7.19).

Figure 5:39 illustrates the estimated marginal means for each post-editing level for visit count instructions, while Figure 5:40 illustrates for visit count UI.

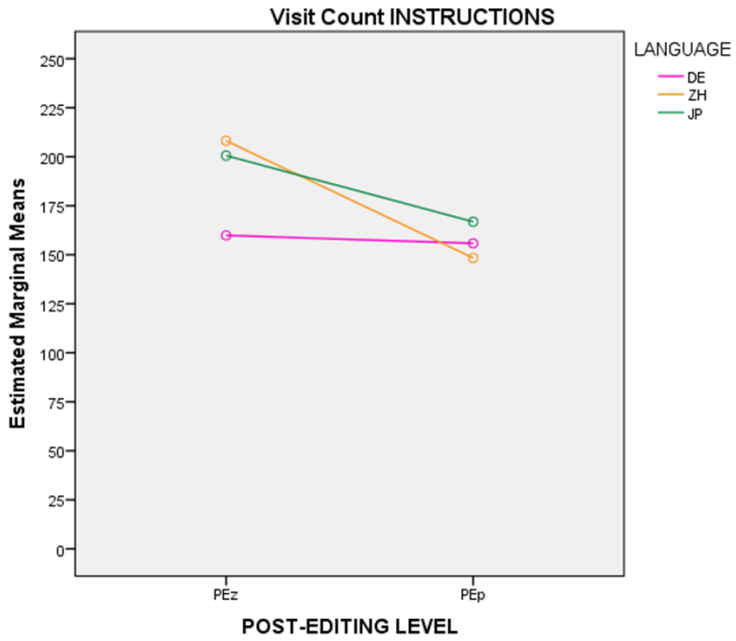


Figure 5:39 - Visit Count Instructions -Translated Content

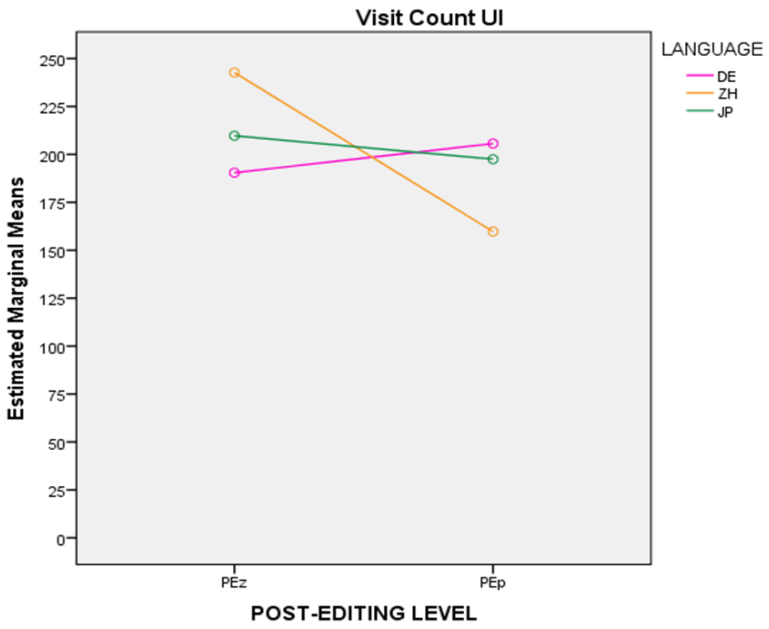


Figure 5:40 - Visit Count UI - Translated Content

When looking at the AOI INST across PE_LEVELs, all PEp groups present fewer visits to the AOI INST when compared to their PEz groups. However, these results are only statistically significant for the Simplified Chinese language (ZH_PEz (M= 208.14, SD=53.36), ZH_PEp (M=148.38, SD=27.89)) at the p <.05 level.

When looking at the AOI UI across PE_LEVELS, both Japanese and the Simplified Chinese PEz groups presents more visits to the AOI UI when compared to the PEp groups. This means that participants of the ZH_PEz and JP_PEz group visited the AOI UI more times than the ZH_PEp and JP_PEp groups. However, this effect was only statistically significant diference between the ZH_PEz group (M=242.71, SD=61.01) and the ZH_PEp group (M=159.75, SD=32.02), at the p<.05 level. Regarding the remaining group DE_PEp, it presents a higher number of visits to the AOI UI when compared to DE_PEz. However, this difference was found not to be statistically significant.

Figure 5:41 shows the differences between VC_INST and VC_UI for each language and post-editing level. When looking at both AOIs (INST *vs* UI) across groups, all groups have more visits in the AOI user interface when compared to the AOI instructions, which means that all groups visited the AOI user interface more times. These results were statistically significant for the groups DE_PEz, DE_PEp,

ZH_PEz, and JP_PEp at the $p<.05$ level; apart from the ZH_PEp and JP_PEz groups which were not statistically significant at the $p>.10$ level.

.

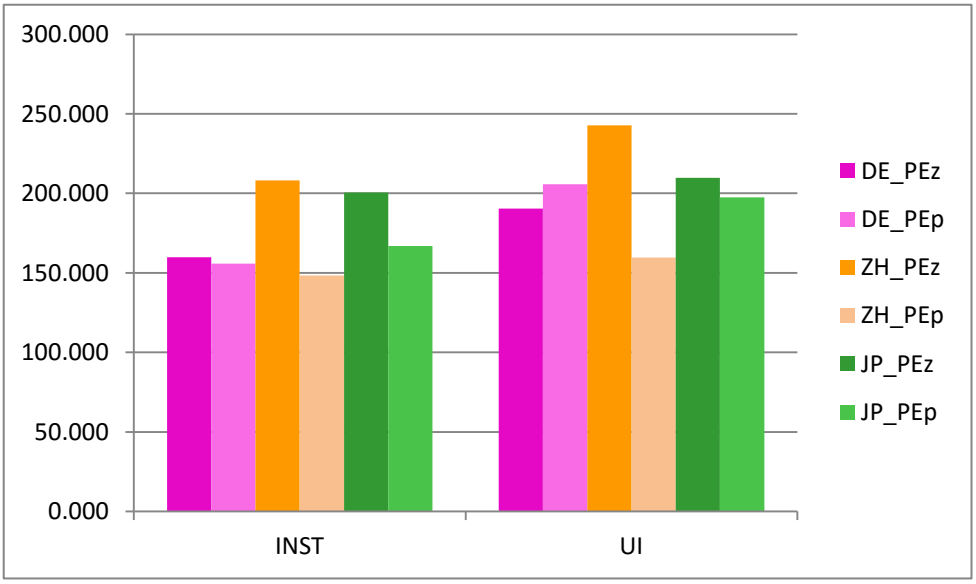


Figure 5:41 - Differences per PE_Level and Language for VC – Translated Content

Comparison with Source

The performance of the participants who used the English Source of the MT Instructions was also computed for visit count via a one-way MANOVA with repeated measures. Table 5:27 shows the mean and standard deviation for each language and their respective post-editing levels (in seconds) for each AOI (instructions and UI) compared to the English Source.

AOIs	Groups		Mean	Std. Deviation
VC_INST	EN	SOURCE	142.14	36.38
	DE	PEz	159.86	33.54
		PEp	155.83	43.89
	ZH	PEz	208.14	53.36
		PEp	148.38	27.89
	JP	PEz	200.57	62.96
		PEp	166.83	46.71
VC_UI	EN	SOURCE	168.29	26.024
	DE	PEz	190.43	60.19
		PEp	205.67	47.39
	ZH	PEz	242.71	61.01
		PEp	159.75	32.02
	JP	PEz	209.71	66.85
		PEp	197.50	67.27

Table 5:27 - Mean and Standard Deviation for Visit Count - Source

The factor PE_LEVEL was found to have a statistically significant effect on VC (F (6, 42) = 2.44, p<.05). The test of within-subjects determined that VC has no significant effect in the interaction with PE_LEVEL (F (6, 42) = .843, p>.10). However, there was a statistically significant difference between VC_INST and VC_UI (F (1, 42) = 23.33, p<.001).

A pairwise comparison found that the participants who used the source instructions (EN (M=155.21, SE=16.72)) had fewer VC when compared to the ZH_PEz (M=225.42, SE=16.72) and JP_PEz (M=205.14, SE=23.65) groups at the p<.05 level.

Figure 5:42 illustrates the estimated marginal means for each language and their PE_LEVEL compared to the Source for the AOI instructions, while Figure 5:43 shows the means for each language and their PE_LEVEL compared to the source for the AOI user interface.

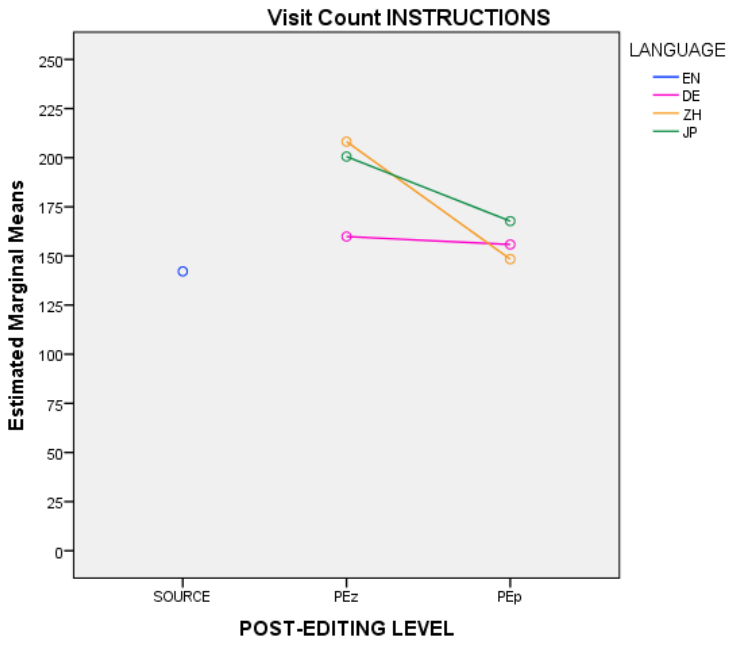


Figure 5:42 - Visit Count Instructions - Source

When looking at the AOI INST across groups (Figure 5:42), the EN_Source group presents fewer visits in the instruction when compared to all the groups. However, these effect was statistically significant only against the ZH_PEz and JP_PEz (p<.05) groups.

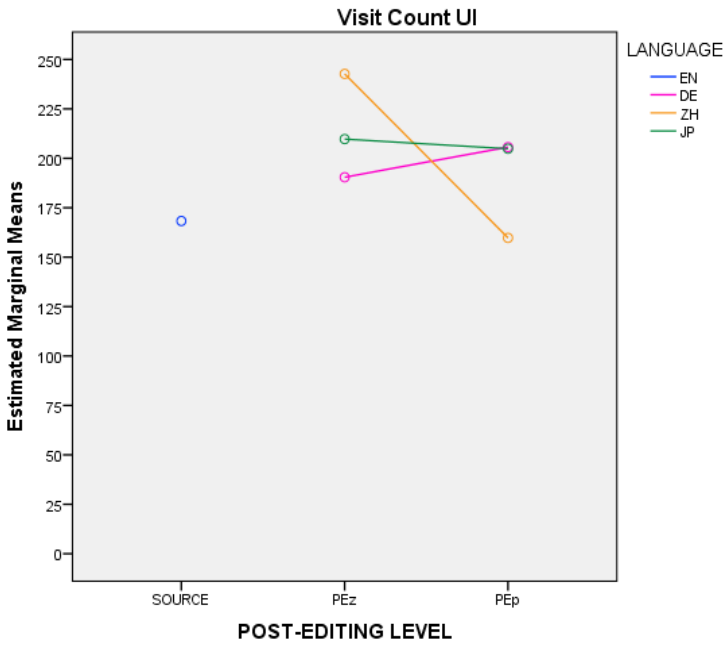


Figure 5:43 - Visit Count UI - Source

When looking at the AOI UI across groups (Figure 5:43), the EN_Source group presents fewer visits in the UI when compared to all the translated language groups. However, this effect was statistically significant only against the ZH_PEz group at the $p<.05$ level.

Figure 5:44 shows the differences between visit counts in the AOI INST and in the AOI UI for each group compared to the EN_Source. The source also follows the previous results in which all the translated language groups have more visits in the AOI user interface when compared to the AOI instructions. This result was statistically significant for the EN_Source group at the $p<.10$ level. This means that for all groups, including the EN_Source group, there was more cognitive effort related to the user interface window against the instruction window.

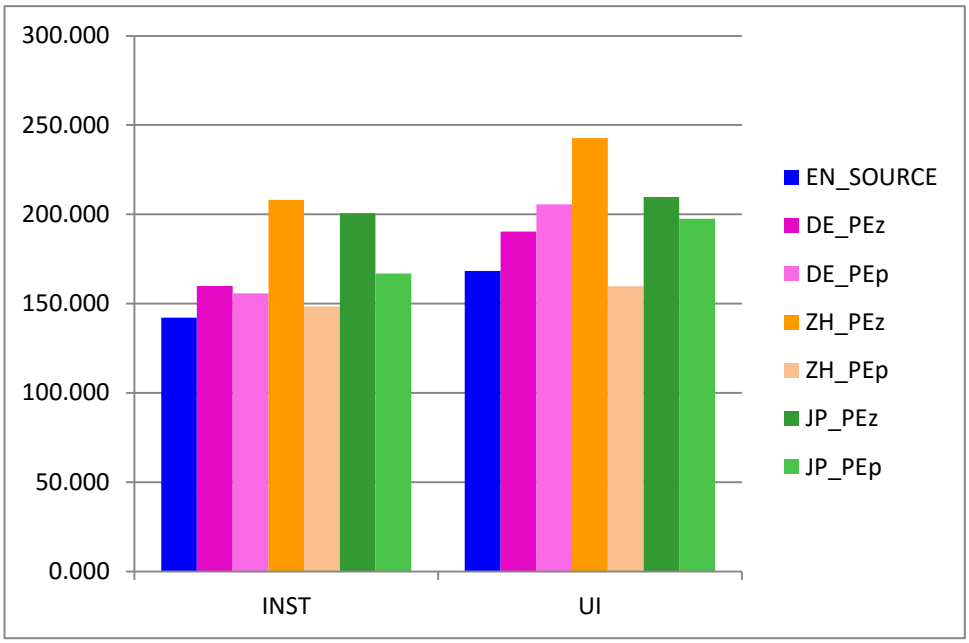


Figure 5:44 - Differences per group for Visit Count - Source

5.1.5.4.2 *HT Instructions*

A two-way MANOVA with repeated measures was conducted in order to compare whether the factors Language and PE_LEVEL have an effect on visit count (VC) for both AOIs: Instruction (VC_INST) and User Interface (VC_UI) for the HT Instructions.

LANGUAGE: The factor Language was found not to have a statistically significant difference on VC, where (F (2, 35) = 1.21, p>.10). This means that when the factor language is considered without distinctions between PE_LEVELs, there is no statistically significant differences across the three translated languages DE (M=29.01, SE=2.64), ZH (M=34.51, SE=2.46), JP (M=30.84, SE=2.64).

POST-EDITING LEVEL: The factor PE_LEVEL was also found not to have a statistically significant difference on VC, where (F (1, 35) = 1.03, p>.10). This means that when the factor PE_LEVEL is considered without distinctions between languages, there is no statistically significant differences across the two post-editing levels PEz (M=32.97, SE=2.07), and PEp (M=29.93, SE=2.14).

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect on VC, where (F (2, 35) = .94, p>.10). This means that the factor language combined with the factor PE_LEVEL do not have a joint effect on VC.

Table 5:28 shows the mean and standard deviation for each language and their respective post-editing levels per AOI (instructions and UI) for the HT instructions.

AOIs	Groups		Mean	Std. Deviation
VC_INST	DE	PEz	28.00	6.32
		PEp	28.83	9.70
	ZH	PEz	37.14	4.53
		PEp	29.63	8.02
	JP	PEz	32.86	11.68
		PEp	28.00	11.30
VC_UI	DE	PEz	28.71	6.58
		PEp	30.50	14.02
	ZH	PEz	40.29	9.91
		PEp	31.00	8.25
	JP	PEz	30.86	9.04
		PEp	31.67	16.91

Table 5:28 - Mean and Standard Deviation for Visit Count - HT Instructions

The test of within-subjects determined that there was not a statistically significant difference between VC_INST and VC_UI (F (1, 35) = 2.33, p>.10) for the HT instructions.

Figure 5:45 illustrates the estimated marginal means for each post-editing level for fixation duration instructions, while Figure 5:46 illustrates for fixation duration UI for the HT Instructions.

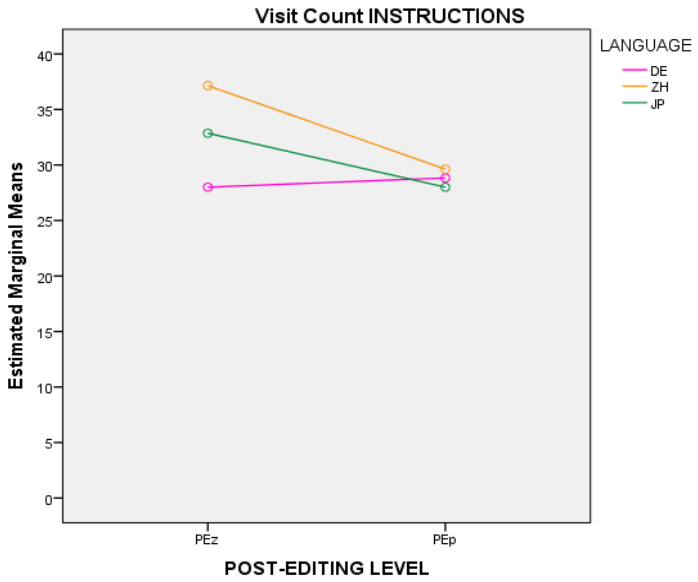


Figure 5:45 - Visit Count Instruction - HT Instructions

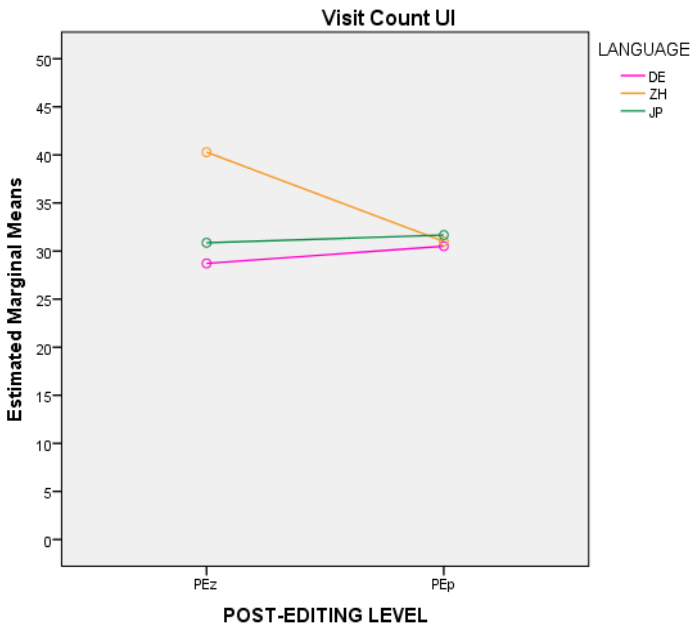


Figure 5:46 - Visit Count UI - HT Instructions

When looking at the AOI INST across PE_LEVELs (Figure 5:45), higher number of visits can be observed for the ZH_PEz and JP_PEz groups when compared to their PEp groups, which indicates that the groups which used the raw machine translated version of the instructions had more cognitive effort observed when reading the instructions. However, these results were not statistically significant (p >.10). The ZH_PEz group, however, shows a moderate statistically significant difference at the p=.11 against the ZH_PEp group. The German language interestingly shows a very close number of visits for both groups for the INST AOI.

When looking at the AOI UI across PE_LEVELS (Figure 5:46), a higher number of visits can be observed for the DE_PEp and JP_PEp groups when compared to their PEz groups. However, neither results were statistically significant (p >.10). The Chinese language, oppositely from the other languages, shows more visits for the PEp group. However, this result showed only a moderate correlation at the p=.11 level.

Figure 5:47 shows the differences between visit count for each language and post-editing level for the instructions and UI area of interest. Although none of the results are statistically significant, the PEp groups from all languages present slightly more visits in the AOI UI than in the AOI INST, following previous results.

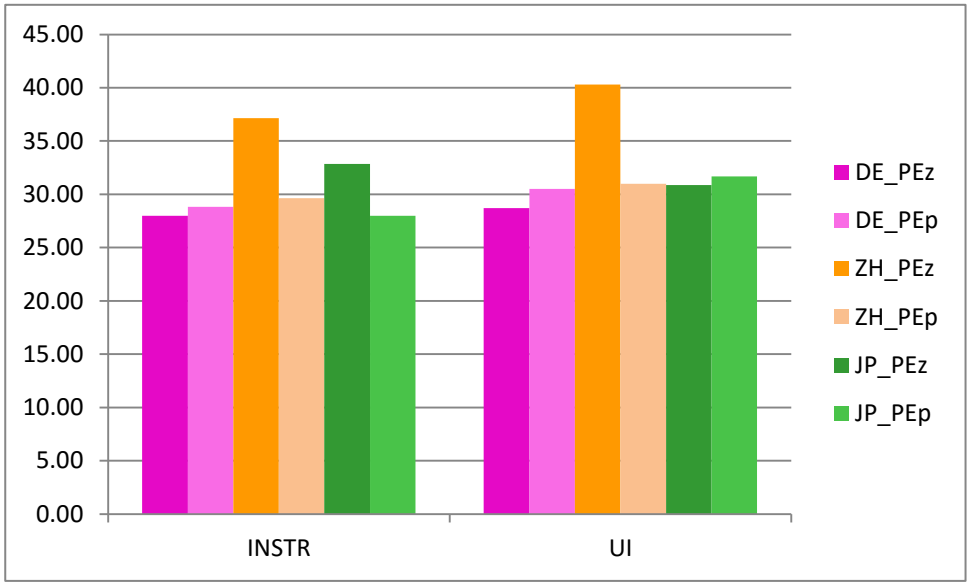


Figure 5:47 - Differences per group for Visit Count - HT Instructions

Comparison with Source

The performance of the participants who used the English Source of the HT Instructions was also computed for visit count via a one-way MANOVA with repeated measures. Table 5:29 shows the mean and standard deviation for each language and their respective post-editing levels (in seconds) for each AOI (instructions and UI) compared to the English Source.

The factor PE_LEVEL was found not to have a statistically significant effect on VC (F (6, 42) = 1.63, p>.10). The test of within-subjects determined that VC had no significant effect in the interaction with PE_LEVEL (F (6, 42) = .781, p>.10). There was a statistically significant difference between VC_INST and VC_UI (F (1, 42) = 2.76, p<.10). A pairwise comparison found that the participants who used the source instructions (EN (M=23.92, SE=3.47)) had fewer VC when compared to the ZH_PEz group (M=38.71, SE=3.47), at the p <.005 level.

AOIs	Groups		Mean	Std. Deviation
VC_INST	EN	SOURCE	23.57	8.04
	DE	PEz	28.00	6.32
		PEp	28.83	9.70
	ZH	PEz	37.14	4.53
		PEp	29.63	8.02
	JP	PEz	32.86	11.68
		PEp	28.00	11.30
VC_UI	EN	SOURCE	24.29	6.97
	DE	PEz	28.71	6.58
		PEp	30.50	14.02
	ZH	PEz	40.29	9.91
		PEp	31.00	8.25
	JP	PEz	30.86	9.04
		PEp	31.67	16.91

Table 5:29 - Mean and Standard Deviation for Visit Count - HT Content - Source

Figure 5:48 illustrates the estimated marginal means for each language and their PE_LEVEL compared to the Source for the AOI instructions, while Figure 5:49 shows the means for each language and their PE_LEVEL compared to the source for the AOI user interface.

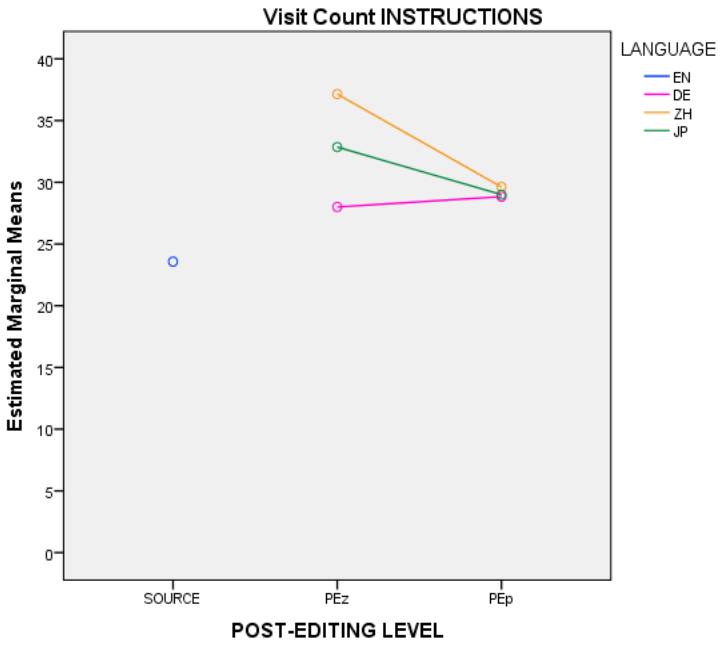


Figure 5:48 - Visit Count Instructions - HT Instructions - Source

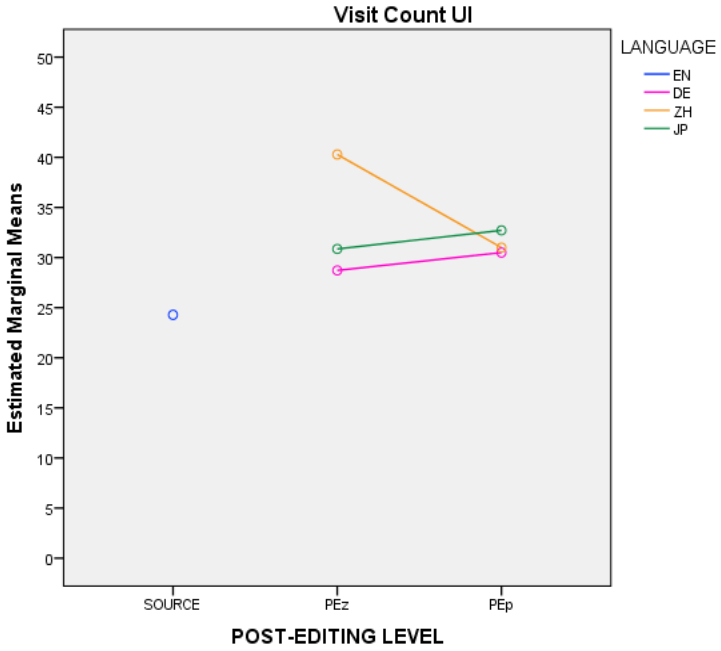


Figure 5:49 - Visit Count UI -HT Instructions - Source

When looking at the AOI INST across groups (Figure 5:48), the EN_Source group presents fewer visits in the instruction when compared to all the groups. However, this effect was statistically significant only against the ZH_PEz (p<.05) and JP_PEz (p<.05) groups. When looking at the AOI UI across groups (Figure 5:49), the EN_Source group presents fewer visits in the UI when compared to all the groups. However, this effect was statistically significant only for the ZH_PEz group at the p<.05 level.

Figure 5:50 shows the differences between VC_INST and VC_UI for each group compared to the EN_Source. The source also follows the previous results in which the PEp groups from all languages present slightly more visits in the AOI UI than in the AOI INST, however, this results was not statistically significant for the EN_Source group at the p>.10 level.

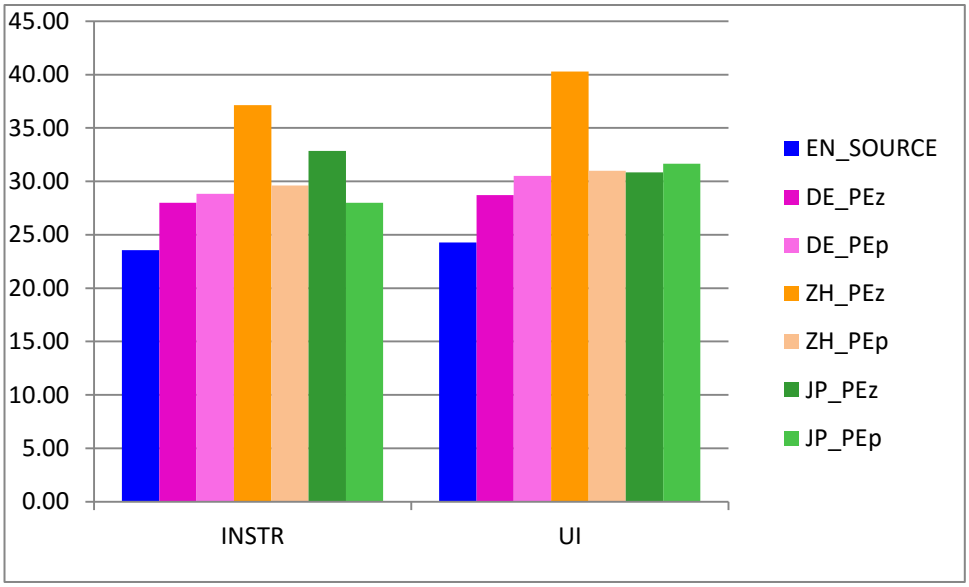


Figure 5:50 - Differences per group for Visit Count – Source – HT instructions

Chapter 6 – Results II

Chapter 5 reported the results for the usability experiments, including the cognitive data. This chapter provides the results for the quality (Section 6.1) and satisfaction (Section 6.2) experiments.

6.1 Quality Experiments

The quality experiments intend to answer the following research questions:

> ***RQ3****: Does the quality evaluation of Post-editing levels PEz and PEp, performed by professional evaluators reflect the results from the empirical usability and satisfaction experiments?*
> ***RQ6****: Does the quality evaluation of the translated languages performed by professional evaluators reflect the results from the empirical usability and satisfaction experiments for Language?*
> ***RQ09****: Does the quality evaluation of the Source Content reflect the results from the empirical usability and satisfaction experiments for Source Content?*

The quality experiments were divided into Source Content and Translated Content. For the Translated Content, the division 'MT Instruction' and 'HT Instructions' was used for the TQA analysis. The TQA for the MT Instruction was calculated via a two-way MANOVA with repeated measures, while the HT Instruction was calculated via a one-way MANOVA with repeated measures[34], where the within-subject factor is defined as PTQ (post-task questionnaire). The comparison MT *vs* HT Instructions was analysed via a two-way MANOVA with repeated measures.

[34] One-way MANOVA is used because only one factor (Language) has more than 2 groups.

6.1.1 Source Content

As described in Section 4.2.2.2, Chapter 4, the source content was analysed with the use of two tools: Source Content Profiler and Coh-Metrix. The aim of this experiment was to understand the features of the selected content (Profiler), as well as the level of comprehension difficulty it presented (Coh-Metrix) in order to understand the profile of the source content. Understanding its profile would, it was assumed, contribute to understanding of the results on acceptability of the translated (raw MT, PEMT and HT) content.

6.1.1.1 SCP

As described in Chapter 4, Section 4.2.4.2, the Source Content Profiler tool allocates a score, the SCP Score, which is taken to be a measure of the quality of a document on a scale from 0 to 100, with a **higher** score indicating higher quality of the document. The results here are reported according to the sub-scores and the final SCP score, as well as the Domain Classification feature score.

SCP score

The SCP profile score for the Online Help content was 60 (out of 100). The results are broken down into:

Average word length:

The average word length for this content was 4 characters per word.

Average sentence length

The average sentence length for this content is 10 words per sentence.

Number of grammar issues;

There were no grammar issues found by the tool.

Number of spelling issues;

There were 51 spelling issues found by the tool. Most were unknown characters such as >> and [].

Number of passive voice issues

There were 4 passive voice sentences in the content.

Percentage of sentences with unusual POS sequences

There were 32 unusual POS sequences, mostly regarding characters such as >> and [].

Domain Classification

The domain detection feature of the SCP tool resulted in:

81% Technical Documentation

7% Training

4% Product

The remaining 8% was classified as Legal 4%, Support 2% and Sales & Marketing 2%.

These results of the Domain Classification feature indicate that the Online Help content contains the terminology and sentence structure expected in instructional content types. Moreover, the results from the SCP tool indicate that the Online Help content used for this research contained few issues and, a score of 60 score indicates, in general, that the English source content was of a good quality.

6.1.1.2 Coh-Metrix

As described in Chapter 4, Section 4.2.4.2, Coh-Metrix is a tool that measures cohesion and coherence for written and spoken texts. The metrics of Coh-Metrix used in this research are Text Easability and Readability.

Text Easability

Syntactic Simplicity – a percentage of 95% was given for syntactic complexity, which indicates that the analysed content presented a high number of sentences which use simpler syntactic structures and have few words, which, in turn, make the text less challenging to read.

Referential Cohesion – a percentage of 33% was given for referential cohesion which indicates that 33% the content analysed contained words and ideas that overlap across sentences and entire text.

Verb Cohesion – a percentage of 0.4% was retrieved for verb cohesion, which indicates that overlapping verbs occurred rarely in the content. The text is likely to include more coherent event structures when verbs are repeated.

Readability

Flesch Reading Ease – A score of 84 (out of 100) was allocated for the analysed content, which indicates the text is relatively easy to read.

The result from the Coh-Metrix tool indicates that, although scores for referential cohesion and verb cohesion are relatively low, in general, the content analysed presents high syntactic simplicity and readability, which can be considered easy to read. It is important to remember that the corpus analysed with the SCP and Coh-Metrix were the excerpts from 6 articles used as instructions for 8 tasks (see Appendix B), not the whole articles. Therefore, it is expected that they might not contain a great overlap of words and verbs as the topics were slightly different from each other. Nonetheless, the choice of measuring the content for referential cohesion and verb cohesion was to identify how much of overlaps those excerpts used as instructions by the participants contained.

6.1.2 Translated Content

As described in Chapter 4, the Translated Content was assessed via a TQA questionnaire sent to the company's moderators. The topics assessed consisted of four main categories:

1- Adequacy

2 -Fluency

3 - Grammar and Syntax

a) Spelling

b) Sentence Structure

4 - Style

a) Terminology

b) Country Standard

As described in Section 4.3.2.2.5 in Chapter 4, the questionnaire was a tailored version from the freely available KantanMT's framework since the industry's partner framework constituted sensitive information. Nonetheless, the questionnaire also had to take into consideration what the industry partner was concerned about. Therefore, categories 1 and 2 were assessed via a 1-4 Likert scales, while categories 3 and 4 via a 1-3 Likert scale (as it was the industry partner's standard). A series of MANOVAs were conducted in order to assess the TQA questionnaire and, because the categories were assessed via different Likert scales, they were divided according to the Likert scale they used. Firstly, fluency and adequacy are calculated together, following, syntax&grammar and style are calculated. Therefore, the results are reported as follows:

1- MT Instructions
   - Adequacy and Fluency
   - Syntax & Grammar and Style

2- HT Instructions
   - Adequacy and Fluency
   - Syntax & Grammar and Style

3- MT instructions *vs* HT Instructions
   - Adequacy and Fluency
   - Syntax & Grammar and Style

MT Instructions are assessed via a two-way MANOVA, while the HT Instructions are assessed via a one-way MANOVA. The comparison MT vs HT Instructions are also calculated via a two-way MANOVA. Therefore, Adequacy and Fluency are the within-subject factor[35] called TQA_1, while Syntax&Grammar and Style are the within-subject factor called TQA_2.

The moderators were presented with the same instructions used by the participants of the usability experiments in the same order (see Table 5:1). Moreover, DE_PEz, DE_PEp, ZH_PEz, ZH_PEp, JP_PEz and JP_PEp refers to the MT

[35] Within-subject factor consists of the dependent variables combined in the MANOVA tests.

Instructions, per language and post-editing level, whereas DE_HT, ZH_HT and JP_HT refers to the HT instructions per language.

6.1.2.1 MT Instructions

Adequacy and Fluency

A two-way MANOVA with repeated measures was conducted in order to compare whether the factors Language and PE_LEVEL have an effect on Adequacy and Fluency (TQA_1).

LANGUAGE: The factor Language was found to have a statistically significant difference on TQA_1, where (F (2, 12) = 3.65, p<.05). This means that when the factor language is considered without distinctions between PE_LEVELs, there is a statistically significant difference across the three translated languages DE (M=2.94, SE=.08), ZH (M=2.79, SE=.08), JP (M=3.09, SE=.08) for the TQA_1.

POST-EDITING LEVEL: The factor PE_LEVEL was found to have a very statistically significant effect on TQA_1, where (F (1, 12) =100.86 p<.001). This means that when the factor PE_LEVEL is considered without distinctions between languages, there is a statistically significant differences across the two post-editing levels PEz (M=2.48, SE=.06), and PEp (M= 3.40, SE=.06).

INTERACTION: The interaction Language*PE_LEVEL was found to have a statistically significant effect TQA_1, where (F (2,12) = 4.92, p<.02). This means that the factor language combined with the factor PE_LEVEL have a joint effect on TQA_1.

Table 6:1 shows the mean and standard deviation for each language and their respective post-editing levels for TQA_1.

Measure TQA_1	Instructions Type		Mean	Std. Deviation
Adequacy	DE	PEz	2.83	0.17
		PEp	3.67	0.17
	ZH	PEz	2.39	0.10
		PEp	3.22	0.25
	JP	PEz	2.78	0.09
		PEp	3.33	0.00
Fluency	DE	PEz	1.72	0.39
		PEp	3.56	0.20
	ZH	PEz	2.50	0.17
		PEp	3.05	0.48
	JP	PEz	2.67	0.17
		PEp	3.61	0.35

Table 6:1 - Mean and Standard Deviation for Adequacy and Fluency - MT Instructions

The test of within-subjects determined that there was a statistically significant difference between Adequacy (M=3.03, SE=.03) and Fluency (M=2.85, SE= .07), where (F (1, 12) = 6.70, p<.05), which means that Adequacy was scored higher when no distinctions between languages and PE_LEVELs are made.

Figure 6:1 illustrates the estimated marginal means for each language and their post-editing level for Adequacy, while Figure 6:2 Illustrates for Fluency.

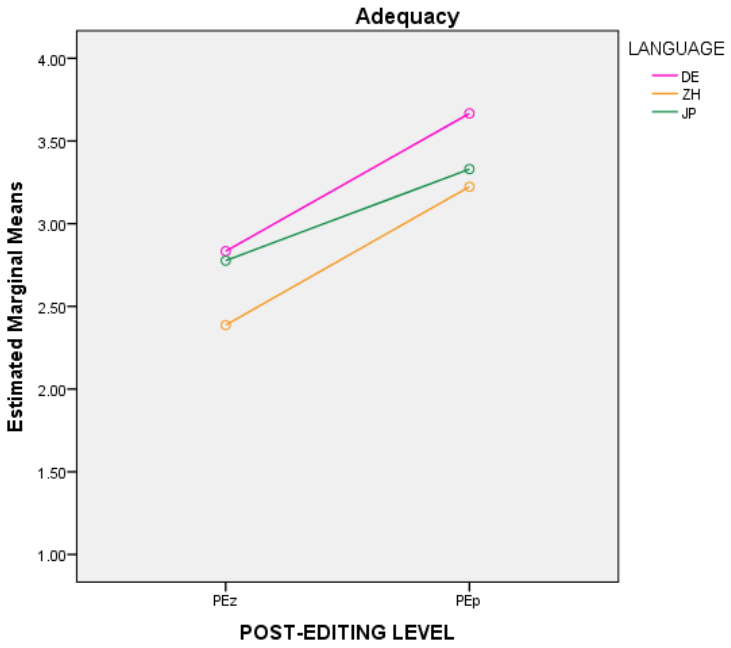


Figure 6:1 - Adequacy - MT Instructions

When looking at Adequacy across PE_LEVELs (Figure 6:1), a higher rating can be observed for DE_PEp instructions when compared to ZH_PEp and JP_PEp

instructions which indicates that the lightly post-edited version of the German language was considered more adequate by the moderators than the post-edited versions of Simplified Chinese and Japanese. This result was statistically significant for both comparisons of DE_PEp against ZH_PEp at the $p<.005$ level and JP_PEp at the $p<.05$ level. When comparing the PEz versions, a slightly higher rating can be observed for the DE_PEz instructions when compared to the JP_PEz, however, this result was not statistically significant. When comparing ZH_PEz against JP_PEz and DE_PEz, a very statistically significant difference was found for both comparisons of ZH_PEz against DE_PEz at the $p < .005$ level and against JP_PEz at the $p < .05$ level, which means that among the raw machine translated instructions, the Simplified Chinese language scored lower for Adequacy. Finally, when comparing the PEp instructions against their own PEz version, DE_PEp, ZH_PEp and JP_PEp show higher rating for Adequacy which indicates that the lightly post-edited version of the instructions were more adequate translation of the source. These results were very statistically significant for the DE_PEp, ZH_PEp and JP_PEp instructions at the $p < .001$ level.

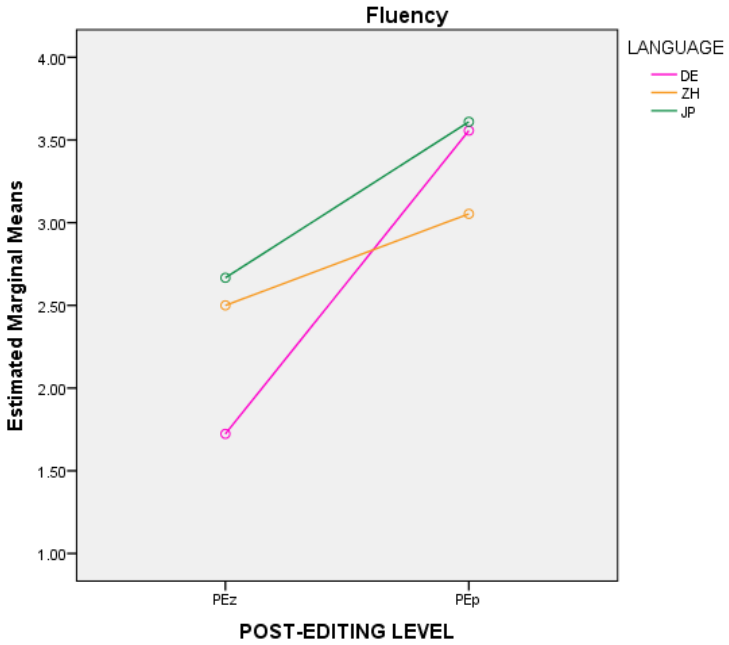


Figure 6:2 - Fluency - MT Instructions

When looking at Fluency across PE_LEVELs (Figure 6:2), a slightly higher rating can be observed for the for JP_PEp instructions when compared to the DE_PEp,

however, this result was not statistically significant. When comparing ZH_PEp against JP_PEp and DE_PEp, a statistically significant difference was found for both comparisons of ZH_PEp against DE_PEp at the $p<.10$ level and against JP_PEz at the $p<.05$ level, which means that among the lightly post-edited instructions, the Simplified Chinese language scored lower for Fluency. When comparing the PEz versions, a higher rating can be observed for JP_PEz instructions when compared to DE_PEz and ZH_PEz instructions, which indicates that the raw machine translated version of the Japanese language was considered to be more fluent by the moderators than the raw machine translated versions of Simplified Chinese and German. However, this result was statistically significant only for the comparison against DE_PEz at the $p<.005$ level. The DE_PEz instruction was also different from the ZH_PEz group at the $p<.05$ level, where the Simplified Chinese instructions scored higher for Fluency. These results indicate that the DE_PEz instructions were considered the least fluent among all instructions from both PE_LEVELs. Finally, when comparing the PEp instructions against their own PEz version, the DE_PEp, ZH_PEp and JP_PEp show higher rating for Fluency which indicates that the lightly post-edited version of the instructions were considered to be more fluent translation of the source. These results were very statistically significant for the DE_PEp at the $p<.001$ level, and for ZH_PEp and JP_PEp instructions at $p<.005$ level.

Syntax&Grammar and Style

A two-way MANOVA with repeated measures was conducted in order to compare whether the factors Language and PE_LEVEL have an effect on Syntax&Grammar and Style (TQA_2).

LANGUAGE: The factor Language was found not to have a statistically significant difference on TQA_2, where (F (2, 12) = .64, $p>.10$). This means that when the factor language is considered without distinctions between PE_LEVELs, there is no statistically significant difference across the three translated languages DE (M=2.38, SE=.10), ZH (M=2.43, SE=.10), JP (M=2.54, SE=.10) for the TQA_2.

POST-EDITING LEVEL: The factor PE_LEVEL was found to have a statistically significant effect on TQA_2, where (F (1, 12) =15.65 p<.005). This means that when the factor PE_LEVEL is considered without distinctions between languages, there is a statistically significant differences across the two post-editing levels PEz (M=2.23, SE=.08), and PEp (M= 2.70, SE=.08).

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect TQA_2, where (F (2,12) = .29, p<.10). This means that the factor language combined with the factor PE_LEVEL do not have a joint effect on TQA_2.

Table 6:2 shows the mean and standard deviation for each language and their respective post-editing levels for TQA_2.

Measure TQA_2	Instructions Type		Mean	Std. Deviation
Spelling	DE	PEz	2.44	0.54
		PEp	2.72	0.25
	ZH	PEz	2.67	0.44
		PEp	2.67	0.44
	JP	PEz	2.39	0.35
		PEp	2.83	0.17
Sentence Structure	DE	PEz	1.11	0.10
		PEp	2.72	0.09
	ZH	PEz	2.22	0.39
		PEp	2.61	0.54
	JP	PEz	1.94	0.10
		PEp	2.50	0.29
Terminology	DE	PEz	2.22	0.25
		PEp	2.39	0.26
	ZH	PEz	1.95	0.63
		PEp	2.67	0.58
	JP	PEz	2.28	0.09
		PEp	2.72	0.25
Country Standards	DE	PEz	2.55	0.25
		PEp	2.89	0.19
	ZH	PEz	2.28	0.19
		PEp	2.72	0.35
	JP	PEz	2.72	0.19
		PEp	3.00	0.00

Table 6:2 - Mean and Standard Deviation for Syntax&Grammar and Style - MT Instructions

The test of within-subjects determined that there was a statistically significant difference between Spelling (M=2.62, SE=.09), Sentence Structure (M=2.28, SE=.07),

Terminology (M=2.37, SE= .09) and Country Standards (M=2.69, SE= .05), where (F (2, 31) = 15.69, p<.001). The results show that all measure where statistically different from each other, apart from Spelling against Country Standards, which means that, when no distinctions between languages and PE_LEVELs are made, Spelling and Country Standards scored higher , followed by Terminology and Sentence Structure respectively .

Figure 6:3 illustrates the estimated marginal means for each language and their post-editing level for Spelling.

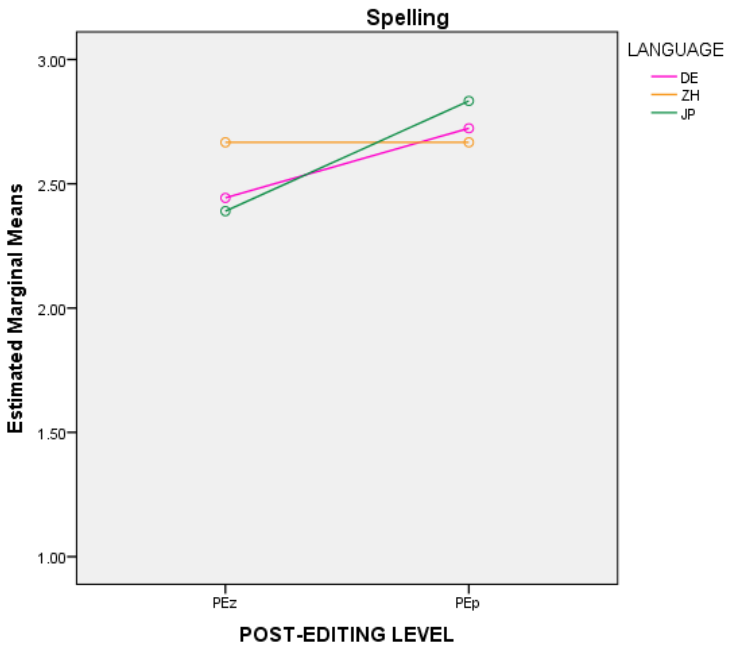


Figure 6:3 - Spelling – MT Instructions

A slightly higher rating for spelling can be observed for JP_PEp instructions when compared to ZH_PEp and DE_PEp instructions; however, this result was not statistically significant for any instruction type. When looking at the PEz versions, a higher rating can be observed for the ZH_PEz instructions when compared to the DE_PEz and JP_PEz, however, this result was not statistically significant for any instruction type. Finally, when comparing the PEp instructions against their own PEz version, the DE_PEp and JP_PEp show higher rating for spelling which indicates that the moderators found fewer spelling issues in the lightly post-edited version of the instructions. However, these results were not statistically significant. The Chinese

language instructions scored the same ratings for both PE_LEVEls, which indicates that both the raw MT and the light PE versions were equally in terms of spelling.

Figure 6:4 illustrates the estimated marginal means for each language and their post-editing level for Sentence Structure.

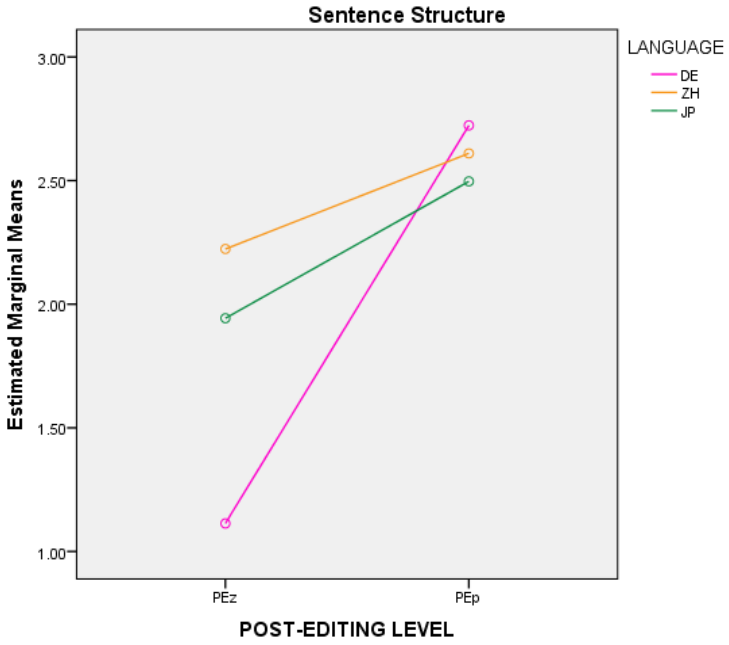


Figure 6:4 - Sentence Structure - MT Instructions

A slightly higher rating for sentence structure can be observed for DE_PEp instructions when compared to ZH_PEp and JP_PEp instructions; however, this result was not statistically significant for any instruction type. When looking at the PEz groups, a higher rating can be observed for the ZH_PEz instructions when compared to the DE_PEz and JP_PEz. This result was statistically significant for the comparisons DE_PEz against the other two PEz instructions at the p <.005 level, which indicates that from the PEz instructions group, the German language contained the highest number of sentence structure issues. Finally, when comparing the PEp instructions against their own PEz version, the DE_PEp, ZH_PEp and JP_PEp show higher ratings for sentence structure which indicates that the lightly post-edited version of the instructions show fewer sentence structure issues. These results were very statistically significant for DE_PEp at the p<.001 level, and for JP_PEp instructions at the p <.05 level. There was no statistically significant difference between the post-editing levels of the Simplified Chinese language.

Figure 6:5 illustrates the estimated marginal means for each language and their post-editing level for Terminology.

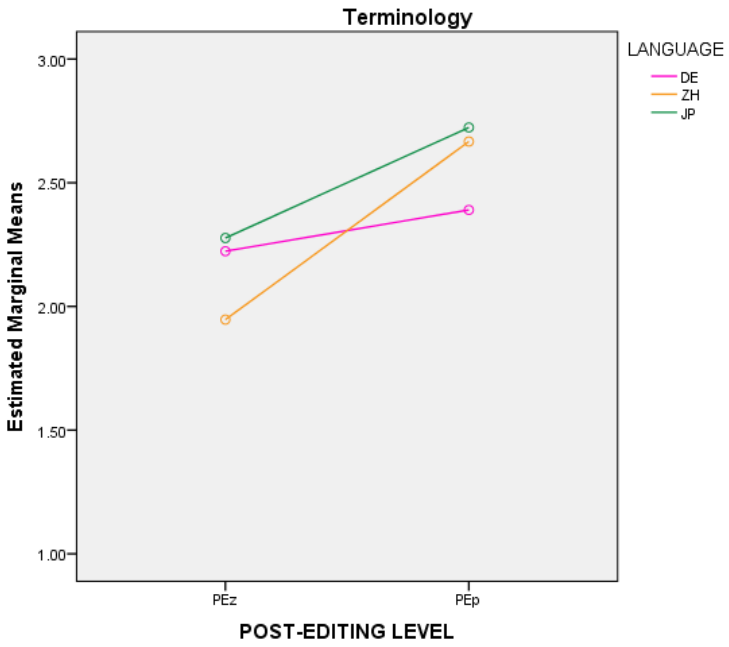


Figure 6:5 - Terminology - MT Instructions

A slightly higher rating for Terminology can be observed for JP_PEp when compared to ZH_PEp and DE_PEp instructions; however, this result was not statistically significant for any of the languages. When looking at the PEz groups, a higher rating can be observed for the JP_PEz instructions when compared to the DE_PEz and ZH_PEz; however, this result was not statistically significant for any languages. When comparing the PEp instructions against their own PEz version, the DE_PEp, ZH_PEp and JP_PEp show higher ratings for Terminology which indicates that the lightly post-edited versions of the instructions have fewer Terminology issues. These results were only statistically significant for the ZH_PEp against the ZH_PEz instructions at the $p<.05$ level. The German and Japanese language did not show statistically significant differences between the PE_LEVELs.

Figure 6:6 illustrates the estimated marginal means for each language and their post-editing level for Country Standards.

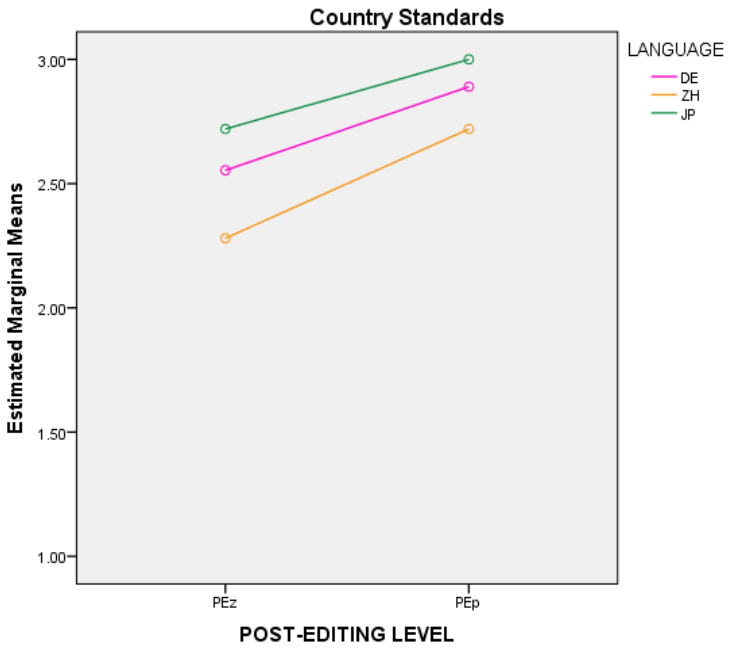


Figure 6:6 - Country Standards - MT Instructions

Similarly to Terminology, the JP_PEp group shows a slightly higher rating for Country Standards when compared to ZH_PEp and DE_PEp instructions; however, this result was not statistically significant for any languages. The same is true for the PEz versions, where a higher rating can be observed for the JP_PEz instructions when compared to the DE_PEz and ZH_PEz; however, this result was only statistically significant for the JP_PEz and ZH_PEz comparison at the $p<.05$ level, which indicates that the Simplified Chinese Language showed more country standards issues than the JP_PEz instructions. When comparing the PEp instructions against their own PEz version, the DE_PEp, ZH_PEp and JP_PEp show higher rating for Country Standards which indicates that the lightly post-edited versions of the instructions show fewer issues. These results were statistically significant for the DE_PEp against the DE_PEz instructions at the $p <.10$ level, and for the ZH_PEp against the ZH_PEz instructions at the $p<.05$ level. The Japanese language did not show statistically significant differences between the PE_LEVELs.

6.1.2.2 HT Instructions

Adequacy and Fluency

The quality of the HT Instructions was also assessed by the moderators and the effects were calculated via a one-way MANOVA.

The factor Language was found not to have a statistically significant difference on TQA_1, where (F (2, 15) = 2.51, p>.10). This means that there is no statistically significant difference across the three translated languages DE (M=3.50, SE=.15), ZH (M=3.04, SE=.15), JP (M=3.41, SE=.15) for the TQA_1 (Adequacy and Fluency). Table 6:3 shows the mean and standard deviation for each language for Adequacy and Fluency (TQA_1).

Measure TQA_1	Instructions Type		Mean	Std. Deviation
Adequacy	DE	HT	3.50	0.55
	ZH	HT	2.92	0.38
	JP	HT	3.33	0.41
Fluency	DE	HT	3.50	0.32
	ZH	HT	3.17	0.61
	JP	HT	3.50	0.55

Table 6:3 - Mean and Standard Deviation for Adequacy and Fluency - HT Instructions

The test of within-subjects determined that TQA_1 did not have any significant effects in the interaction with Language (F (2, 15) = .28, p>.10). There was not a statistically significant difference between Adequacy and Fluency (F (1, 15) = 1.0, p>.10). Figure 6:7 illustrates the estimated marginal means for each language.

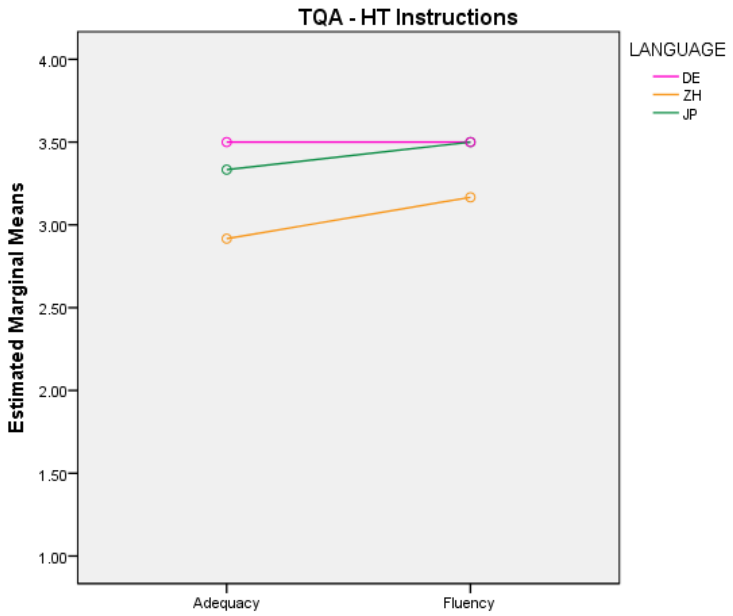


Figure 6:7 - Adequacy and Fluency - HT Instructions

When looking at Adequacy across languages, a slightly higher rating can be observed for DE_HT when compared to JP_HT and ZH_HT instructions; however, this result was only statistically significant for the DE_HT against the ZH_HT instructions at the $p<.05$ level, which indicates that the human translated instructions for Chinese were not considered as adequate as the HT for German and Japanese.

When looking at Fluency, a higher rating can be observed for the JP_HT and DE_HT instructions when compared to the ZH_HT, however, this result was only statistically, which indicates that all languages showed similar level of Fluency for the HT instruction.

Syntax&Grammar and Style

A one-way MANOVA was conducted in order to identify the effect of language on TQA_2 (Spelling, Sentence Structure, Terminology and Country Standards) for the HT instructions.

The factor Language was found to have a statistically significant difference on TQA_2, where (F (2, 15) = 2.88, $p<.10$). This means that there is a statistically significant difference across the three translated languages DE (M=2.68, SE=.10), ZH

(M=2.56, SE=.10), JP (M=2.91, SE=.10) for the TQA_2. Table 6:4 shows the mean and standard deviation for each language for TQA_2, while Figure 6:8 illustrates the estimated marginal means for each language.

Measure TQA_2	Instructions Type		Mean	Std. Deviation
Spelling	DE	HT	2.58	0.49
	ZH	HT	2.83	0.41
	JP	HT	2.92	0.20
Sentence Structure	DE	HT	2.5	0.32
	ZH	HT	2.42	0.58
	JP	HT	2.75	0.27
Terminology	DE	HT	2.92	0.20
	ZH	HT	2.42	0.38
	JP	HT	3.00	0.00
Country Standards	DE	HT	2.75	0.42
	ZH	HT	2.58	0.49
	JP	HT	3.00	0.00

Table 6:4 - Mean and Standard Deviation for Syntax&Grammar and Style - HT Instructions

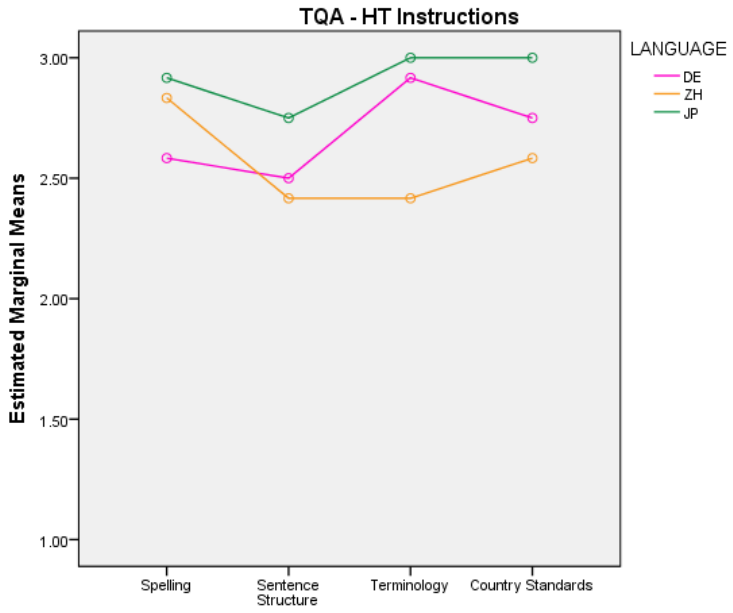


Figure 6:8 - TQA_2 - HT Instructions

When looking at Spelling across languages, a slightly higher rating can be observed for JP_HT when compared to ZH_HT and DE_HT instructions; however, this result was not significant. For Sentence Structure, JP_HT presents a higher rating when compared to DE_HT and ZH_HT instructions; however, this result was not significant as well. For the Terminology measure, JP_HT and DE_HT present a

higher rating when compared to ZH_HT instructions. These results were statistically significant for both comparisons of ZH_HT against DE_HT and JP_HT at the p <.005 level, which indicates that the human translated instructions of Simplified Chinese presented issues regarding Terminology. Finally, when looking at Country Standards a higher rating can be observed for the JP_HT instructions when compared to the ZH_HT and DE_HT instructions, however, this result was only statistically significant for the JP_HT against ZH_HT comparison at the p <.10 level, which indicates that the human translation of Simplified Chinese presented issues with Country Standards that were significantly different from the Japanese language.

6.1.2.3 MT Instructions vs HT Instructions

Adequacy and Fluency

In order to identify whether there are differences between the ratings for the HT instructions and ratings for the MT Instructions (PEz and PEp) for TQA_1 (Adequacy and Fluency), a two-way MANOVA was conducted. Table 6:5 shows the mean and standard deviation, for each language and Instruction type for TQA_1, while Figure 6:9 illustrates the estimated marginal means for each language and PE_LEVEL for Adequacy.

Measure TQA_1	Instruction Type		Mean	Std. Deviation
Adequacy	DE	MT	2.83	0.17
		PE	3.67	0.17
		HT	3.50	0.55
	ZH	MT	2.39	0.10
		PE	3.22	0.25
		HT	2.92	0.38
	JP	MT	2.78	0.09
		PE	3.33	0.00
		HT	3.33	0.41
Fluency	DE	MT	1.72	0.39
		PE	3.56	0.20
		HT	3.50	0.32
	ZH	MT	2.50	0.17
		PE	3.05	0.48
		HT	3.17	0.61
	JP	MT	2.67	0.17
		PE	3.61	0.35
		HT	3.50	0.55

Table 6:5 - Mean and Standard Deviation for Adequacy and Fluency - MT *vs* HT Instructions

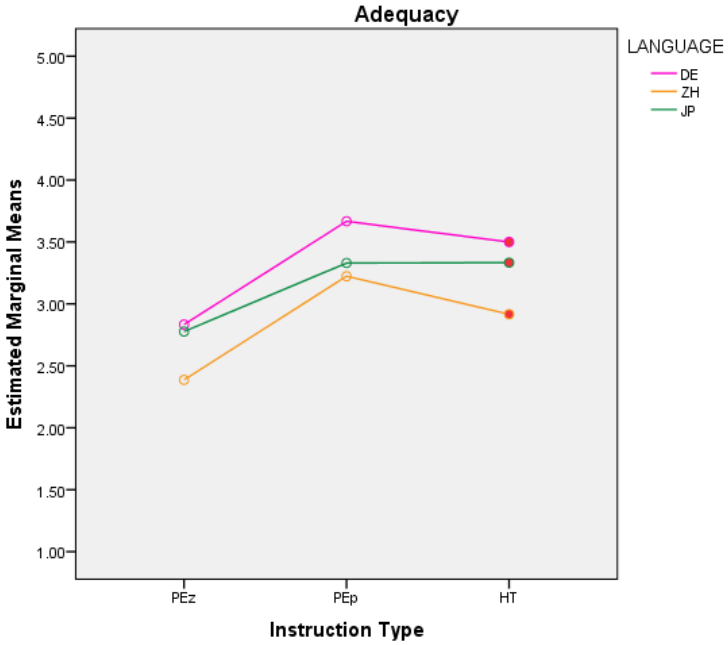


Figure 6:9 - Adequacy - MT *vs* HT Instructions

When comparing Adequacy for the HT Instructions with MT Instructions for the German language, a slightly higher rating for the PEp Instructions can be observed, but no statistically significant differences between the PEp and HT instruction types. When comparing the PEz and HT instructions for German, a statistically significant difference was found at the $p<.05$ level. The results for German mean that the PEp and HT Instructions were comparatively adequate, but the PEz instructions showed Adequacy issues.

The Japanese language shows similar ranking for the HT and PEp instructions regarding Adequacy. When comparing the PEz and HT instructions, a strong statistically significant difference was found at the $p<.05$ level. These results for Japanese indicate that the PEz instructions presented Adequacy issues, but PEp and HT instructions were comparatively adequate.

Finally, for the Simplified Chinese language, a higher rating for the PEp Instructions can be observed when compared to the HT, but no statistically significant differences between the two instruction types were found. When comparing the PEz and HT instructions, a strong statistically significant difference was found at the $p<.05$ level. The results for Chinese follow the previous results for German and Japanese, where the PEz instructions show Adequacy issues, being statistically different from both PEp and HT, but PEp and HT instructions are

comparatively adequate. Figure 6:10 illustrates the estimated marginal means for each language and PE_LEVEL for Fluency.

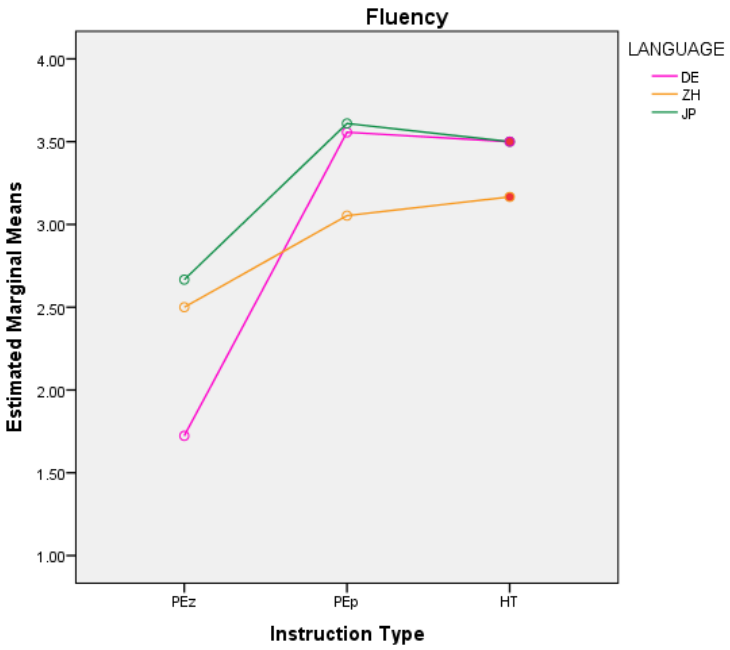


Figure 6:10 - Fluency - MT *vs* HT Instructions

When comparing Fluency for the HT Instructions with MT Instructions for the German language, a slightly higher rating for the PEp Instructions can be observed, but no statistically significant differences between the two instruction types. When comparing the PEz and HT instructions for German, a very statistically significant difference was found at the $p<.001$ level. The results for German mean that the PEp and HT Instructions were comparatively fluent, but the PEz instructions showed Fluency issues.

The Japanese language shows a slightly higher rating for the HT Instructions when compared to the HT and PEp instructions, but no statistically significant differences between the two instruction types. When comparing the PEz and HT instructions, a statistically significant difference was found at the $p<.05$ level. These results for Japanese mean that the PEz instructions presented Fluency issues but the PEp and HT instructions were comparatively fluent.

Finally, for the Simplified Chinese language, a higher rating for the PEp Instructions can be observed when compared to the HT, but no statistically significant differences between the two instruction types were found. When

comparing the PEz and HT instructions, a statistically significant difference was found at the $p<.05$ level. The results for Chinese follow the previous results for German and Japanese, where the PEz instructions show Fluency issues, being statistically different from both PEp and HT, but PEp and HT instructions are comparatively fluent.

Syntax&Grammar and Style

In order to identify whether there are differences between the ratings for the HT instructions and ratings for the MT Instructions (PEz and PEp) for Syntax&Grammar and Style (TQA_2), a two-way MANOVA was conducted. Table 6:6 shows the mean and standard deviation, for each language and Instruction type for TQA_2, while Figure 6:11 illustrates the estimated marginal means for each language and PE_LEVEL for Spelling.

Measure TQA_2	Instruction Type		Mean	Std. Deviation
Spelling	DE	MT	2.44	0.54
		PE	2.72	0.25
		HT	2.58	0.49
	ZH	MT	2.67	0.44
		PE	2.67	0.44
		HT	2.83	0.41
	JP	MT	2.39	0.35
		PE	2.83	0.17
		HT	2.92	0.20
Sentence Structure	DE	MT	1.11	0.10
		PE	2.72	0.09
		HT	2.50	0.32
	ZH	MT	2.22	0.39
		PE	2.61	0.54
		HT	2.42	0.58
	JP	MT	1.94	0.10
		PE	2.50	0.29
		HT	2.75	0.27
Terminology	DE	MT	2.22	0.25
		PE	2.39	0.26
		HT	2.92	0.20
	ZH	MT	1.95	0.63
		PE	2.67	0.58
		HT	2.42	0.38
	JP	MT	2.28	0.09
		PE	2.72	0.25
		HT	3.00	0.00
Country Standards	DE	MT	2.55	0.25
		PE	2.89	0.19
		HT	2.75	0.42
	ZH	MT	2.28	0.19
		PE	2.72	0.35
		HT	2.58	0.49
	JP	MT	2.72	0.19
		PE	3.00	0.00
		HT	3.00	0.00

Table 6:6 - Mean and Standard Deviation for Syntax&Grammar and Style - MT *vs* HT Instructions

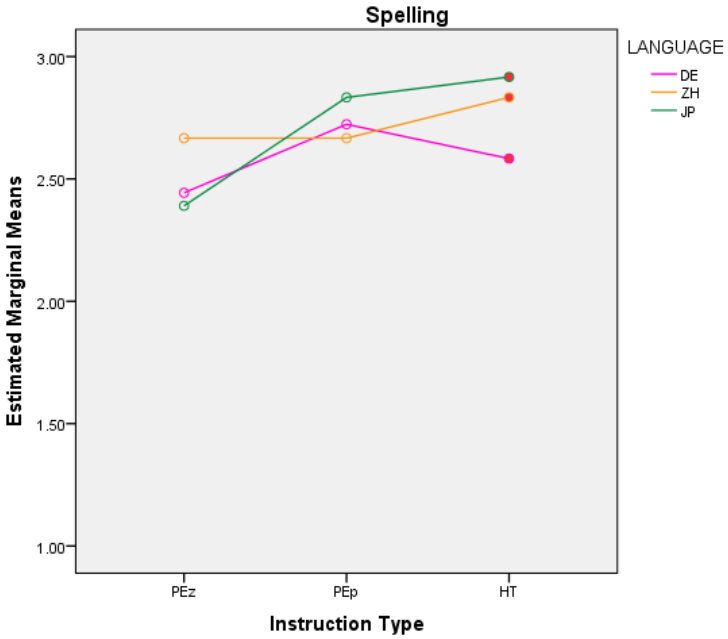


Figure 6:11 - Spelling - MT *vs* HT Instructions

The Japanese language shows a slightly higher ranking for the HT instructions when compared to the PEp regarding Spelling, but no statistically significant difference was found. When comparing the PEz and HT instructions, a statistically significant difference was found at the $p<.10$ level. These results for Japanese indicates that the PEz instructions presented Spelling issues, but PEp and HT instructions were comparatively correct.

For the Simplified Chinese language, a higher rating for the HT Instructions can be observed when compared to the PEp and PEz instructions, but no statistically significant differences between the three instruction types were found.

Finally, when comparing the HT Instructions with PEp Instructions for the German language, a slightly higher rating for the PEp Instructions can be observed, but no statistically significant differences between the two instruction types. When comparing the PEz and HT instructions for German, no statistically significant difference was as well. The results for German mean that the PEp, PEz and HT Instructions were comparatively correct regarding spelling issues.

Figure 6:12 illustrates the estimated marginal means for each language and PE_LEVEL for Sentence Structure.

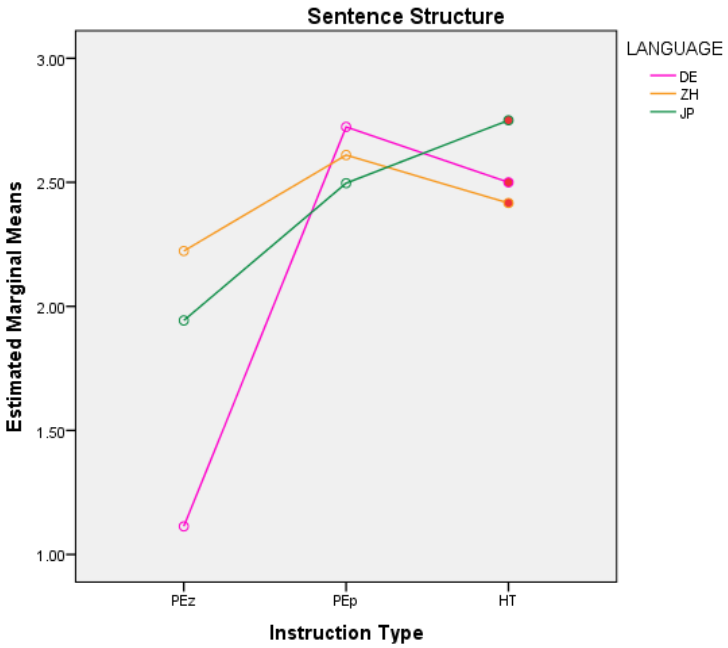


Figure 6:12 - Sentence Structure - MT *vs* HT Instructions

When comparing Sentence Structure for the HT Instructions with PEp Instructions for the German language, a higher rating for the PEp Instructions can be observed, but no statistically significant differences between the two instruction types. When comparing the PEz and HT instructions, a very statistically significant difference was found at the $p<.001$ level. The results for German mean that the PEp and HT Instructions were comparatively structured, but the PEz instructions showed structure issues.

The Japanese language shows a higher rating for the HT Instructions when compared to the PEp instructions, but no statistically significant differences between the two instruction types. When comparing the PEz and HT instructions, a statistically significant difference was found at the $p<.10$ level. These results for Japanese indicate that the PEz instructions presented structure issues, but PEp and HT instructions were comparatively well structured.

Finally, for the Simplified Chinese language, a higher rating for the PEp Instructions can be observed when compared to the HT, but no statistically significant differences between the two instruction types were found. When comparing the PEz and HT instructions, again, no statistically significant differences

between the two instruction types were found. The results for Chinese indicate that the three instruction types were comparatively structured.

Figure 6:13 illustrates the estimated marginal means for each language and PE_LEVEL for Terminology.

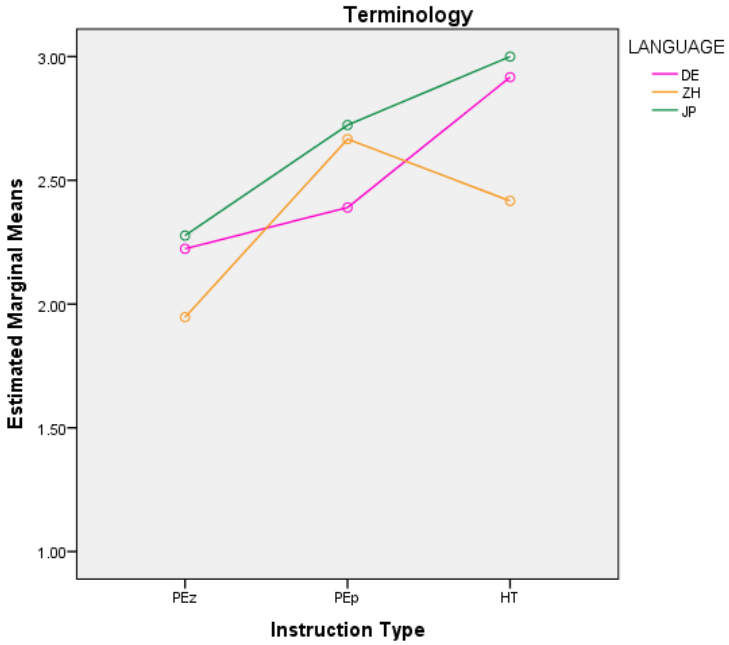


Figure 6:13 - Terminology - MT *vs* HT Instruction

Regarding Terminology, the Japanese language shows a higher ranking for the HT instructions when compared to the PEp, but no statistically significant difference was found between the two instruction types. When comparing the PEz and HT instructions, a statistically significant difference was found at the $p<.005$ level. These results for Japanese indicates that the PEz instructions presented Terminology issues, but PEp and HT instructions were comparatively correct.

For the Simplified Chinese language, a higher rating for the PEp Instructions can be observed when compared to the HT, but no statistically significant differences between them were found. When comparing the PEz and HT instructions, a statistically significant difference was found at the $p<.05$ level. The results for Chinese indicate that the PEz instructions show Terminology issues, being statistically different from both PEp and HT, but PEp and HT instructions are comparatively correct regarding terminology.

For the German language, the HT instructions presents a higher rating for Terminology compared to both PEp and PEz instructions. This results was

statistically significant between DE_HT and DE_PEp at the *p* <.05 level, and between DE_HT and DE_PEz at the *p* <. 005 level. The results for German mean that the HT Instructions were better in terminology when compared to both raw machine translated and post-edited instructions.

Figure 6:14 illustrates the estimated marginal means for each language and PE_LEVEL for Country Standards.

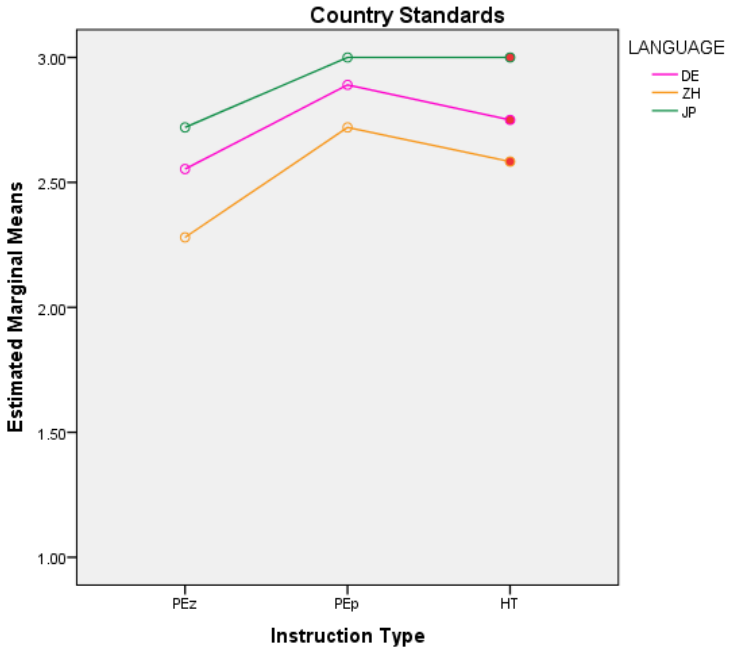


Figure 6:14 - Country Standards - MT *vs* HT Instruction

The Japanese language shows similar ranking for the HT and PEp instructions regarding Country Standards. When comparing the PEz and HT instructions, no statistically significant differences between the three instruction types were found.

For the Simplified Chinese language, a higher rating for the PEp Instructions can be observed when compared to the HT but no statistically significant differences between the three instruction types were found. No statistically significant differences were found for the comparison HT and PEz for Chinese.

Finally, the German language follows the results of Simplified Chinese, where a higher rating for the PEp Instructions can be observed against HT but no statistically significant differences found. The same for the comparison HT and PEz of German.

The results for German mean that the PEp, PEz were comparatively correct compared to the HT instructions regarding country standards issues.

6.2 Satisfaction Experiments

The satisfaction experiments intend to answer the following research questions:

> ***RQ2****: Does Post-editing Level have an effect on satisfaction?*
>
> ***RQ5****: How do different target languages compare in terms of satisfaction for both PEp and PEz content?*
>
> ***RQ8:*** *How does satisfaction with Source Content compare with satisfaction with translated content (PEp and PEz)?*

Three experiments were implemented for Satisfaction: The post-task questionnaire (6.2.1) displayed after the usability experiments; Moderators' ratings for satisfaction (6.2.2), displayed together with the quality TQA questionnaire; and the web survey satisfaction (6.2.3). The post-task questionnaire was analysed via a two-way MANOVA with repeated measures. For the Moderators' rating for satisfaction, the distinction 'MT Instructions' and 'HT' Instructions' was used. The MT Instructions was calculated via a two-way ANOVA, while the HT Instruction was calculated via a one-way ANOVA. The comparison MT *vs* HT Instructions was analysed via a two-way ANOVA.

6.2.1 Post-task Questionnaire

A two-way MANOVA with repeated measures was conducted in order to compare whether the factors Language and PE_LEVEL have an effect on the statements in the post-task questionnaire (PTQ).

LANGUAGE: The factor Language was found not to have a statistically significant difference on PTQ, where (F (2, 56) = 1.43, p>.10). This means that when the factor language is considered without distinctions between PE_LEVELs, there is no statistically significant differences across the three translated languages DE (M=2.38, SE=.16), ZH (M=2.6, SE=.13), JP (M=2.70, SE=.11).

POST-EDITING LEVEL: The factor PE_LEVEL was found to have a statistically significant effect on PTQ, where (F (1, 56) = 4.59, p<.05). This means that when the

factor PE_LEVEL is considered without distinctions between languages, there is a statistically significant differences across the two post-editing levels PEz (M=2.42, SE=.108), and PEp (M=2.75, SE=.11).

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect PTQ, where (F (2, 56) = .15, p>.10). This means that the factor language combined with the factor PE_LEVEL do not have a joint effect on PTQ.

Table 6:7 shows the mean and standard deviation for each language and their respective post-editing levels for each statement[36].

[36] In order to be able to compute a two-way MANOVA with all the statements, statements 5 and 8 were revised to 1- strongly agree to 5-strongly disagree. Therefore, lower rates for these two statements indicate more agreement with them.

Statements	Groups		Mean	Std. Deviation
1- The instructions were usable.	DE	PEz	2.63	1.19
		PEp	3.17	0.75
	ZH	PEz	2.40	1.07
		PEp	3.70	0.82
	JP	PEz	2.54	1.05
		PEp	3.07	1.22
2 - The instructions were comprehensible.	DE	PEz	2.13	0.99
		PEp	3.00	1.41
	ZH	PEz	2.60	0.84
		PEp	3.40	0.84
	JP	PEz	2.69	1.11
		PEp	2.80	1.08
3 - The instructions allowed me to complete all of the necessary tasks	DE	PEz	1.88	0.99
		PEp	3.00	1.55
	ZH	PEz	2.20	1.23
		PEp	3.00	1.25
	JP	PEz	2.08	1.04
		PEp	2.40	0.83
4 - I was satisfied with the instructions provided.	DE	PEz	1.63	0.52
		PEp	2.00	1.10
	ZH	PEz	2.20	1.03
		PEp	2.50	0.97
	JP	PEz	2.08	0.95
		PEp	2.67	1.35
5 - The instructions could be improved upon.	DE	PEz	1.13	0.35
		PEp	1.33	0.82
	ZH	PEz	1.50	0.53
		PEp	2.00	1.15
	JP	PEz	2.00	1.08
		PEp	1.93	0.96
6 - I would consult these instructions again in the future	DE	PEz	3.38	1.30
		PEp	3.17	0.98
	ZH	PEz	3.20	1.40
		PEp	3.70	0.67
	JP	PEz	3.00	1.22
		PEp	3.60	1.06
7 - I would be able to use the software again in the future without re-reading the instructions.	DE	PEz	2.50	1.07
		PEp	2.83	0.98
	ZH	PEz	3.10	1.37
		PEp	1.00	0.00
	JP	PEz	2.46	1.20
		PEp	2.80	1.21
8 - I would rather have seen the source (English) version of the instructions	DE	PEz	1.63	0.74
		PEp	2.67	1.63
	ZH	PEz	2.00	0.67
		PEp	3.20	1.40
	JP	PEz	3.00	1.08
		PEp	3.13	0.83
9 - I would recommend the software to a friend or a colleague	DE	PEz	2.63	0.74
		PEp	2.33	1.21
	ZH	PEz	3.40	1.26
		PEp	3.10	0.74
	JP	PEz	3.46	0.97
		PEp	3.00	1.00

Table 6:7 - Mean and Standard Deviation for Satisfaction Post-task Questionnaire - Translated Content

The test of within-subjects determined that there was a statistically significant difference among the PTQ statements, where (F (6, 351) = 16.27, p<.001).

Figure 6:15 illustrates the estimated marginal means for each translated language and their post-editing level for statement 1.

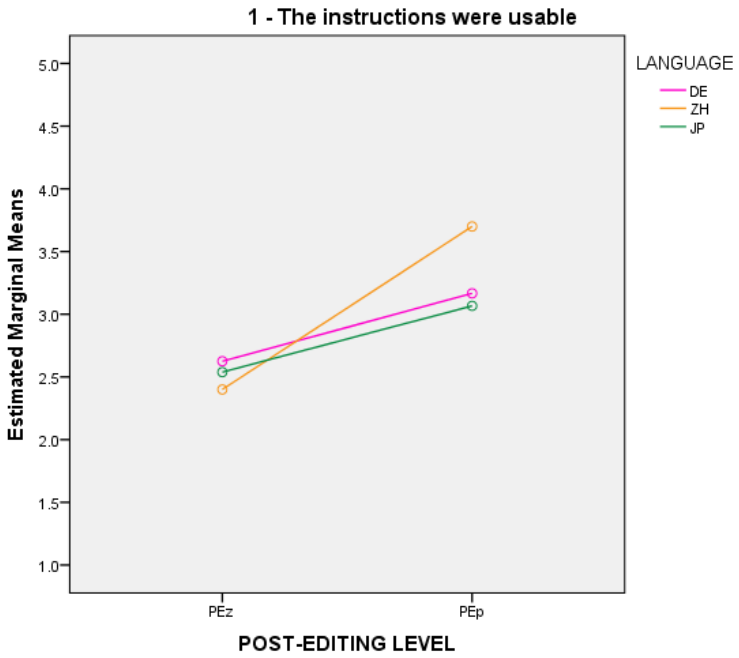


Figure 6:15 - Statement 1 - Translated Content

When looking at statement 1 across PE_LEVELs, the ZH_PEp group presents higher ratings when compared to the DE_PEp and JP_PEp groups, however, there were no statistically significant differences for the PE_LEVEL PEp across the translated languages. This means that participants from all the PEp groups (DE_PEp, ZH_PEp and JP_PEp) found the instructions usable to a similar extent. Regarding the PE_LEVEL PEz, the ZH_PEz presents lower ratings when compared to the DE_PEz and JP_PEz. This result was not statistically significant for any of the PEz groups.

When comparing PEp groups against their PEz, the DE_PEp, ZH_PEp and JP_PEp show higher ratings for statement 1 which indicates that the groups which used the lightly post-edited translated version of the instructions considered the instructions more usable. These results were statistically significant only for the ZH_PEp (M=3.7, SE=.33) when compared to ZH_PEz (M=2.4, SE=.33) at the p <.05.

Figure 6:16 illustrates the estimated marginal means for each translated language and their post-editing level for statement 2.

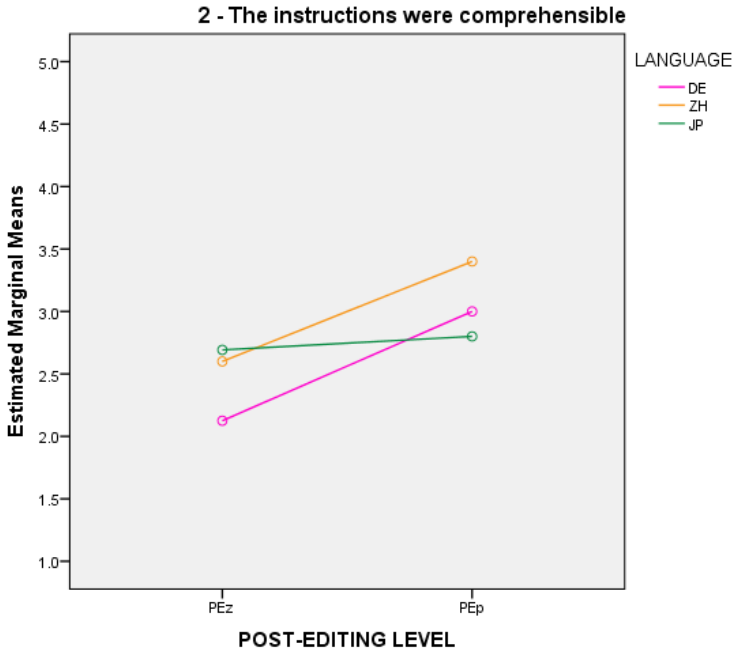


Figure 6:16 - Statement 2 - Translated Content

Similarly to the previous statement, a higher rating for statement 2 can be observed for ZH_PEp group when compared to all the other PEp groups. However, this result was not statistically significant for any of the groups (p>.10). Regarding the PEz groups, the German group shows the lowest ratings for comprehensibility compared to ZH_PEz and JP-PEz. However, these results were not statistically significant (p>.10).

When comparing PEp groups against their PEz groups, the DE_PEp, ZH_PEp and JP_PEp show higher ratings for statement 2 which indicates that the groups which used the lightly post-edited translated version of the instructions considered the instructions more comprehensible. These results were statistically significant only for the ZH_PEp (M=3.4, SE=.33) when compared to ZH_PEz (M=2.6, SE=.33) at the p <.10 level.

Figure 6:17 illustrates the estimated marginal means for each translated language and their post-editing level for statement 3.

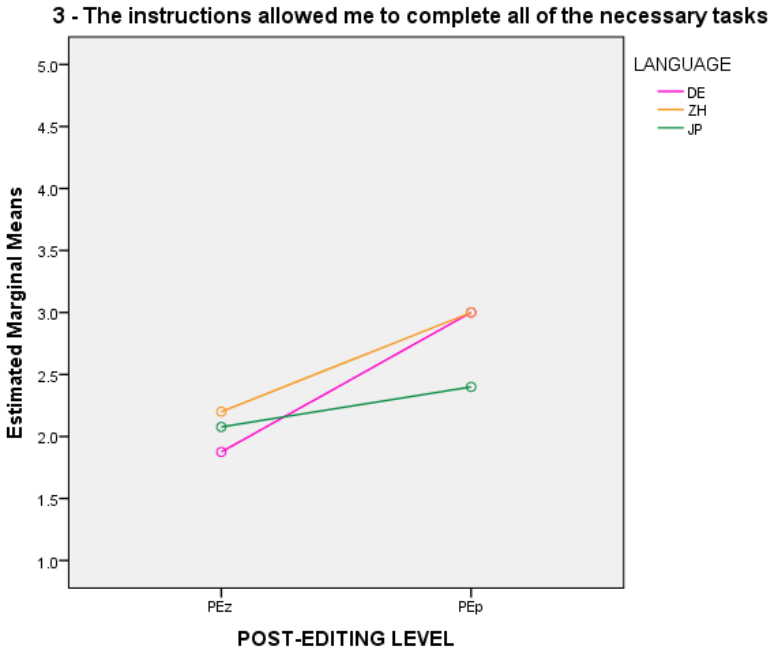


Figure 6:17 - Statement 3 - Translated Content

A higher rating for statement 3 can be observed for DE_PEp and ZH_PEp groups when compared to the other JP_PEp group. However, this result was not statistically significant for any of the groups (p >.10). No statistically significant differences were found among the PEz levels (p >.10).

When comparing PEp groups against their PEz groups, the DE_PEp, ZH_PEp and JP_PEp show higher rating for statement 3 which indicates that the groups which used the lightly post-edited translated version of the instructions considered the instructions more helpful when performing the tasks. These results were statistically significant only for the DE_PEp (M=3.0, SE=.45) when compared to DE_PEz (M=1.87, SE=.39) at the p <.05 level. There was also a moderate statistically significant difference between ZH_PEp (M=3.0, SE=.33) when compared to ZH_PEz (M=2.2, SE=.33) at the p =.11 level.

Figure 6:17 illustrates the estimated marginal means for each translated language and their post-editing level for statement 4.

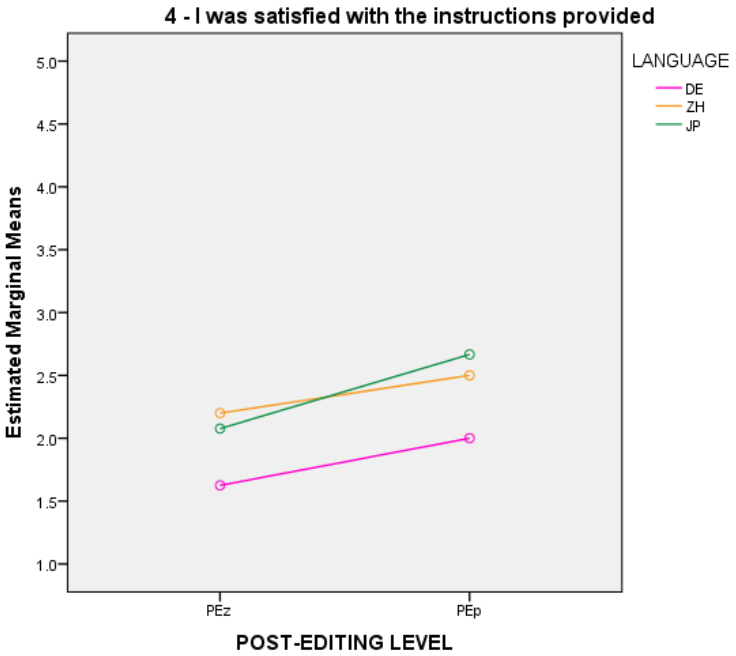


Figure 6:18 - Statement 4 - Translated Content

A higher rating for statement 4 can be observed for JP_PEp group when compared to the DE_PEp and ZH_PEp groups. However, this result was not statistically significant for any of the groups (p >.10). When comparing PEp groups against their PEz, the DE_PEp, ZH_PEp and JP_PEp show higher rating for statement 4 which indicates that the groups which used the lightly post-edited translated version of the instructions were more satisfied with the version they used. However, these results were not statistically significant for any groups (p>.10).

Figure 6:19 illustrates the estimated marginal means for each translated language and their post-editing level for statement 5. Note that for statement 5, a low rating indicates that participant considered the instruction needed more improvements.

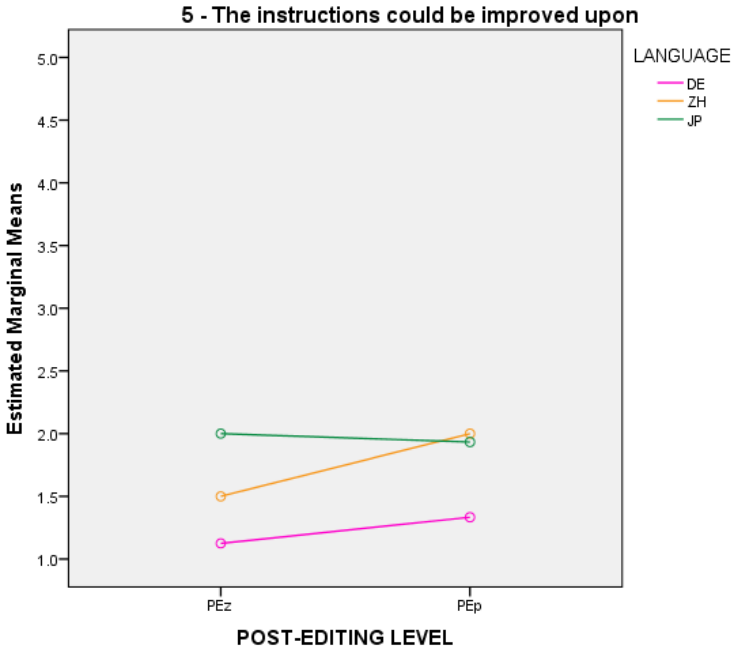


Figure 6:19 - Statement 5 - Translated Content

A *lower* rating for statement 5 can be observed for the German language compared to Japanese and Simplified Chinese languages (no distinction between PE_LEVELS). This means that participants who used the German instructions (raw machine translation and post-edited) considered that the instructions needed more improvement than participants from the other languages. This result was statistically significant at the $p<.05$ level for DE (M=1.22, SE=.24) when compared to the JP language (M=1.96, SE=.17).

DE_PEp and ZH_PEp groups show higher ratings for statement 5 when compared to the DE_PEz and ZH_PEz groups, which indicates that participants who used the PEp version considered that the instructions needed less improvement than their PEz groups. However, none of the results were statistically significant at the $p>.10$ level. Interestingly, the JP_PEz group shows slightly higher ratings for statement 5 when compared to their PEp group. However, no statistically significant difference was found for this result.

Figure 6:20 illustrates the estimated marginal means for each translated language and their post-editing level for statement 6.

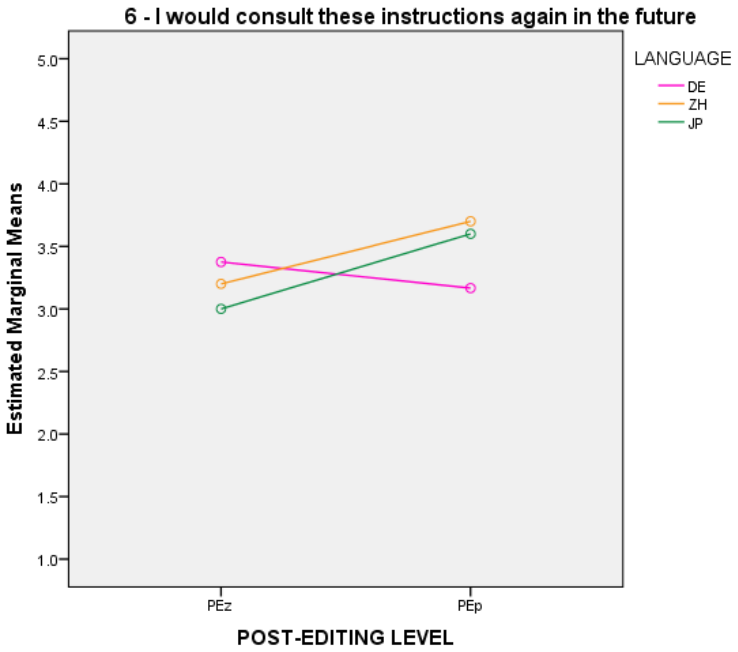


Figure 6:20 - Statement 6 - Translated Content

A higher rating for statement 6 can be observed for ZH_PEp and JP_PEp groups when compared to the DE_PEp group. When looking at the PEz groups, DE_PEz shows a higher rating when compared to JP_PEz and ZH_PEz. However, these results were not statistically significant for any of the groups (p >.10). When comparing PEp groups against their PEz, ZH_PEp and JP_PEp show higher rating for statement 6 which indicates that the groups which used the lightly post-edited translated version of the instructions were more inclined to consult the instructions again. DE_PEp, however, shows *lower* ratings for statement 6 when compared to the DE_PEz groups, which indicates that for the German language, the group who used the raw machine translated version was more inclined to consult the instructions again. However, none of these results were statistically significant at the p >.10 level.

Figure 6:21 illustrates the estimated marginal means for each translated language and their post-editing level for statement 7.

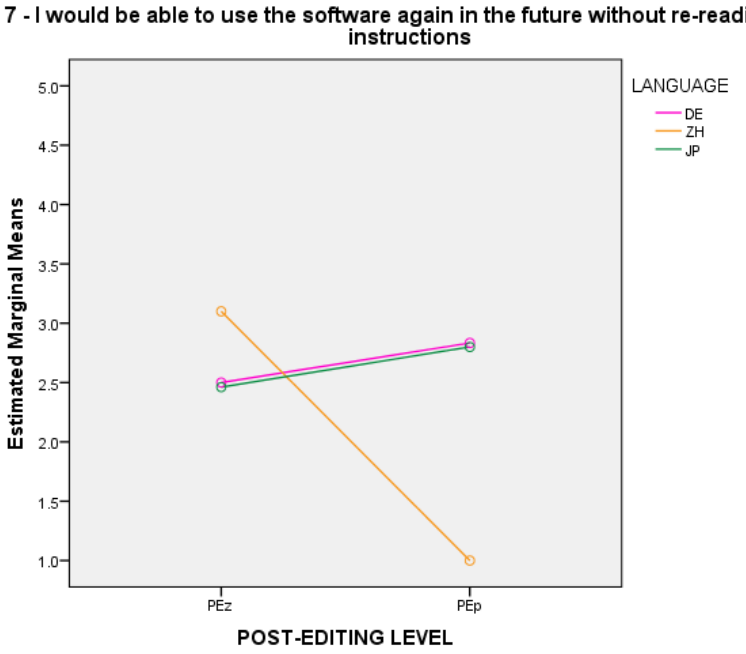


Figure 6:21 - Statement 7 - Translated Content

A higher rating for statement 7 can be observed for the ZH_PEz group compared to DE_PEz and JP_PEz groups. When looking at the PEp groups, DE_PEp and JP_PEp show a higher rating when compared to the ZH_PEp group. These results were statistically significant for both DE_PEp (M=2.83, SE=.44) and JP_PEp (M=2.80, SE=.28) when compared to the ZH_PEp (M=1.00, SE=.34) group, which shows that the Chinese group considered that they would need the instructions again to reuse the software.

DE_PEp and JP_PEp groups show higher ratings for statement 7 when compared to the DE_PEz and JP_PEz groups. However, none of the results were statistically significant at the p >.10 level. Interestingly, the ZH_PEz group shows higher ratings for statement 7 when compared to their PEp group. This result was very statistically significant at the p<.005 level, which shows that for the Chinese language, participants who used the post-edited version were more inclined to rely on the instructions again to re-use the software.

Figure 6:22illustrates the estimated marginal means for each translated language and their post-editing level for statement 8. Note that for statement 8, a low rating indicates that participant would have preferred to see the English version of the instructions instead of the translated version they used.

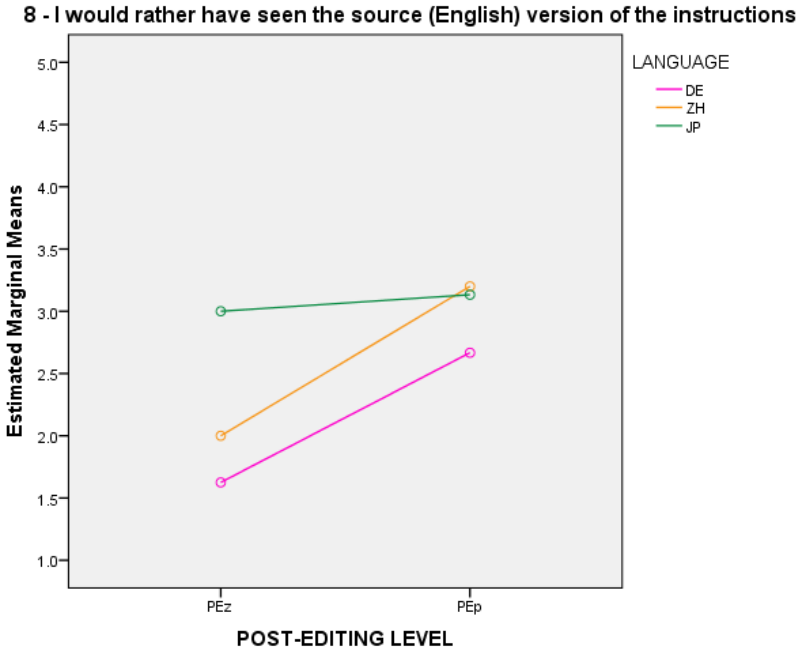


Figure 6:22 - Statement 8 - Translated Content

A *lower* rating for statement 8 can be observed for the German language compared to Japanese and Simplified Chinese languages (no distinction between PE_LEVELS). This means that participants who used the German instructions (raw machine translation and post-edited) would prefer to use the English version more so than participants from the other languages. This result was statistically significant at the $p<.05$ level for DE (M=2.14, SE=.28) when compared to the JP language (M=3.06, SE=.20).

All the 3 PEp groups (DE_PEp, ZH_PEp and JP_PEp) show higher ratings for statement 8 when compared to the DE_PEz, ZH_PEz and JP_PEz groups, which indicates that participants who used the raw machine translated version were more inclined to use the English version than the PEp participants. These results were statistically significant for the ZH_PEp (M=3.2, SE=.33) when compared to the ZH_PEz (M=2.0, SE=.33) group ($p<.05$) and DE_PEp (M=2.66, SE=.43) groups when compared to the DE_PEz (M=1.62, SE=.37) group ($p<.10$). The JP_PEp group shows slightly higher ratings for statement 8 when compared to their PEz group, however, this result was not statistically significant

Finally, Figure 6:23 illustrates the estimated marginal means for each translated language and their post-editing level for statement 9.

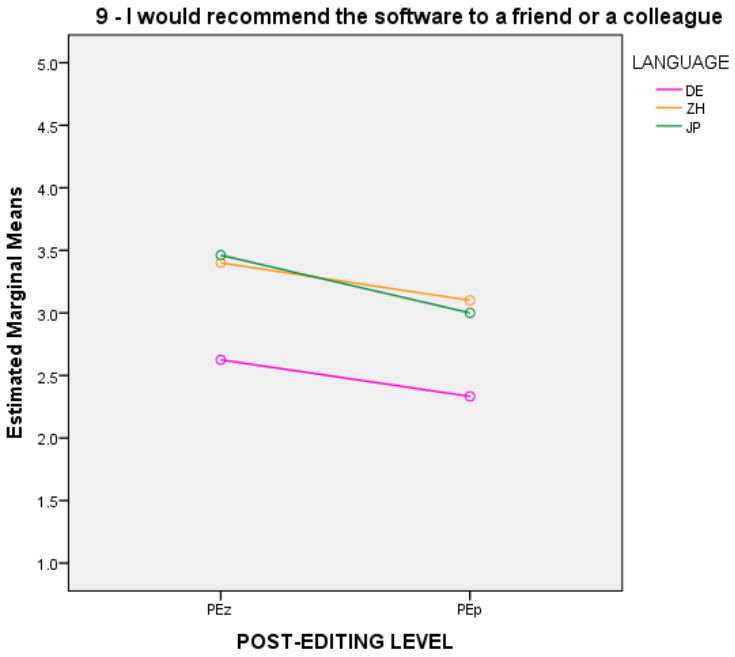


Figure 6:23 - Statement 9 - Translated Content

A *lower* rating for statement 9 is found for the German language compared to Japanese and Simplified Chinese languages (no distinction between PE_LEVELS). This means that participants who used the German instructions (raw machine translation and post-edited) were less inclined to recommend the software used to perform the tasks. This result was statistically significant at the $p<.05$ level for DE (M=2.47, SE=.26.) when compared to the JP language (M=3.23, SE=.18.) and Simplified Chinese (M=3.25, SE=.22).

All the PEz groups rated higher for this statement when compared to their PEp groups. However, none of these results were statistically significant at the $p>.10$ level.

Comparison with Source

The pot-task questionnaire was also displayed for participants who used the English Source instructions to perform the usability tasks, and so, satisfaction was also computed via a one-way MANOVA with repeated measures. Table 6:8 shows the mean and standard deviation for the English Source. Note that participants of the English group did not see statement 8 (I would rather have seen the source (English) version of the instructions).

Statements	Groups		Mean	Std. Deviation
1- The instructions were usable.	EN	SOURCE	3.75	0.71
2 - The instructions were comprehensible.	EN	SOURCE	3.50	0.76
3 - The instructions allowed me to complete all of the necessary tasks	EN	SOURCE	2.63	0.74
4 - I was satisfied with the instructions provided.	EN	SOURCE	2.75	0.71
5 - The instructions could be improved upon.	EN	SOURCE	1.63	0.52
6 - I would consult these instructions again in the future	EN	SOURCE	3.50	0.76
7 - I would be able to use the software again in the future without re-reading the instructions.	EN	SOURCE	2.88	1.13
9 - I would recommend the software to a friend or a colleague	EN	SOURCE	3.38	0.92

Table 6:8 - Mean and Standard Deviation PTQ - Source

The factor PE_LEVEL was found to have a statistically significant effect on PTQ (F (6, 63) = 1.46, p>.10). The test of within-subjects determined that PTQ (when all statements are considered) has a very significant effect in the interaction with PE_LEVEL (F (38, 401) = 3.68, p<.001). There was also a statistically significant difference between the PTQ statements (F (6, 401) = 22.30, p<.001).

A pairwise comparison found that the participants who used the source instructions (EN (M=2.66, SE=0.20)) presented higher ratings for all the PTQ statements when compared to the DE_PEz (M=2.16, SE=0.20) at the p<.10 level.

Figure 6:24 illustrates the estimated marginal means for each translated language and their post-editing level compared to the EN_Source for statement 1. A higher rating for statement 1 can be observed for the EN_Source (M=3.75, SE=.36) group when compared to all the groups. However, this effect was statistically significant only when compared to the PEz groups from all languages, DE_PEz (M=2.62, SE=.36), ZH_PEz (M=2.40, SE=.32) and JP_PEz (M=2.53, SE=.28) at the p<.05 level. This means that the participants who used the English version of the instructions considered the instructions more usable when compared to participants who used the PEz versions. There was no statistically significant difference between the EN_Source groups and the PEp groups, which indicates that participants who used the lightly post-edited version of the instructions considered them as usable as the EN_Source participants.

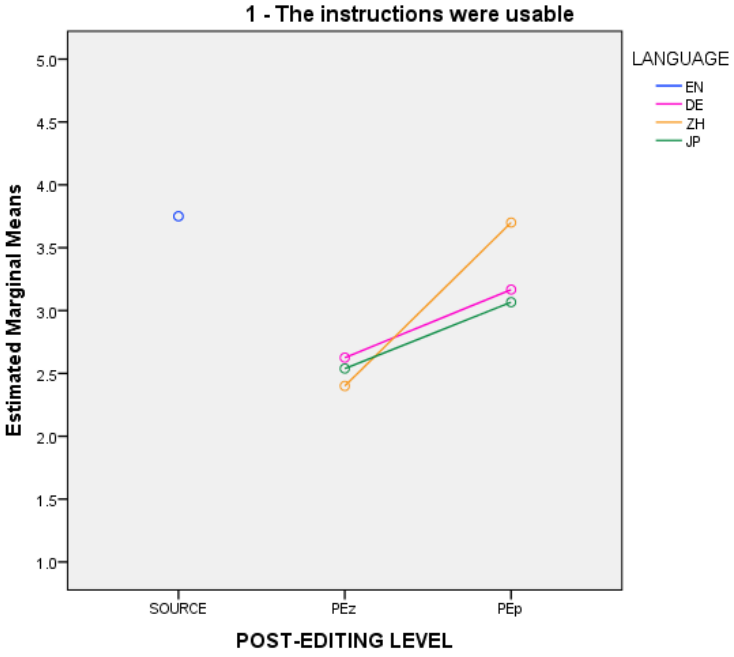


Figure 6:24 - Statement 1 - Source

Figure 6:25 illustrates the estimated marginal means for each translated language and their post-editing level compared to the EN_Source for statement 2.

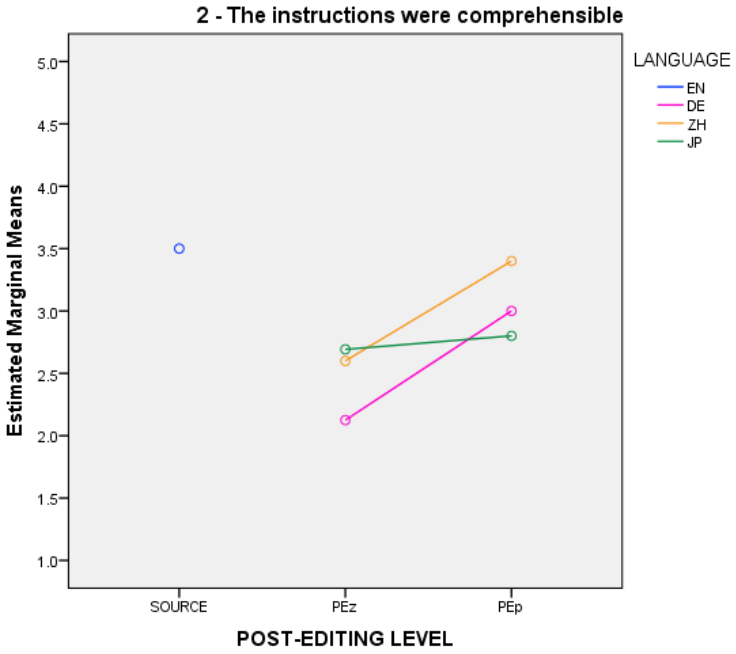


Figure 6:25 - Statement 2 - Source

Similarly to the previous rating for statement 1, for the statement 2, the EN_Source presents a higher rating (M=3.5, SE=.35) when compared to all the groups. This effect was statistically significant only when compared to the PEz groups from all languages, DE_PEz (M=2.12, SE=.35.) at the $p<.05$ level, ZH_PEz (M=2.6, SE=.32) and JP_PEz (M=2.69, SE=.28) at the $p<.10$ level. This means that the participants who used the English version of the instructions considered the instructions more comprehensible when compared to participants who used the PEz versions. There was no statistically significant difference between the EN_Source groups and the PEp groups, which indicates that participants who used the lightly post-edited version of the instructions considered them as comprehensible as the EN_Source participants.

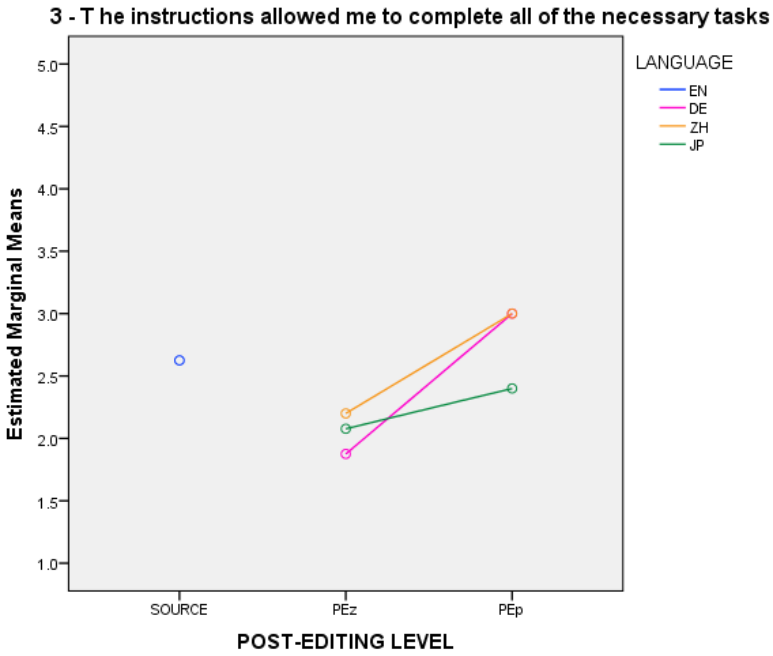


Figure 6:26 - Statement 3 - Source

Figure 6:26 illustrates the estimated marginal means for each translated language and their post-editing level compared to the EN_Source for statement 3. For the statement 3, the EN_Source presents a higher rating when compared to all the PEz groups and the JP_PEp group. Interestingly, the DE_PEp and ZH_PEp groups show higher rating than the EN_Source group. This indicates that participants who used the DE and ZH lightly post-edited versions of the instructions for languages considered them to be more helpful when performing the tasks than the participants who used

the source instructions. However, these results were not statistically significant (p>.10).

Figure 6:27 illustrates the estimated marginal means for each translated language and their post-editing level compared to the EN_Source for statement 4. The EN_Source (M=2.75, SE=.36) presents a higher rating for statement 4 when compared to all PEz and PEp groups, which indicates that EN_Source participants were more satisfied with the instructions. However, this result was statistically significant only for the DE_PEz (M=1.6, SE=.36) group. It is worth mentioning that both ZH_PEp (M=2.5, SE=.32) and JP_PEp (M=2.6, SE=.26) groups have very close ratings when compared to the EN_Source, and because of the lack of a statistically significant difference among the three groups, we can conclude that participants from these three groups were similarly satisfied with the instructions.

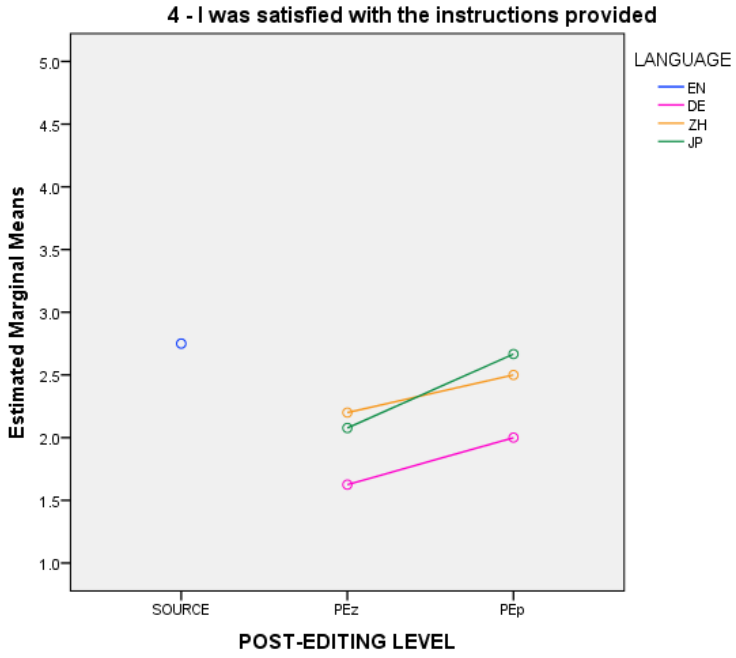


Figure 6:27 - Statement 4 - Source

Figure 6:28 illustrates the estimated marginal means for each translated language and their post-editing level compared to the EN_Source for statement 5. Note that a lower rating demonstrates higher agreement with the statement.

The EN_Source presents a higher rating when compared to all the DE_PEz, DE_PEp and ZH_PEz groups, which indicates that the EN_Source group disagree with the statement more than these other groups. The JP_PEp, JP_PEz and ZH_PEp groups present higher ratings when compared to the EN_Source group, which indicates that

these three translated content groups consider that the instructions were less in need of improvement. However, none of these comparisons were statistically significant (p>.10). It is worth mentioning, however, that for all groups, the rating means for statement 5 are relatively low, which means all participants judged that the instructions needed improvement.

.

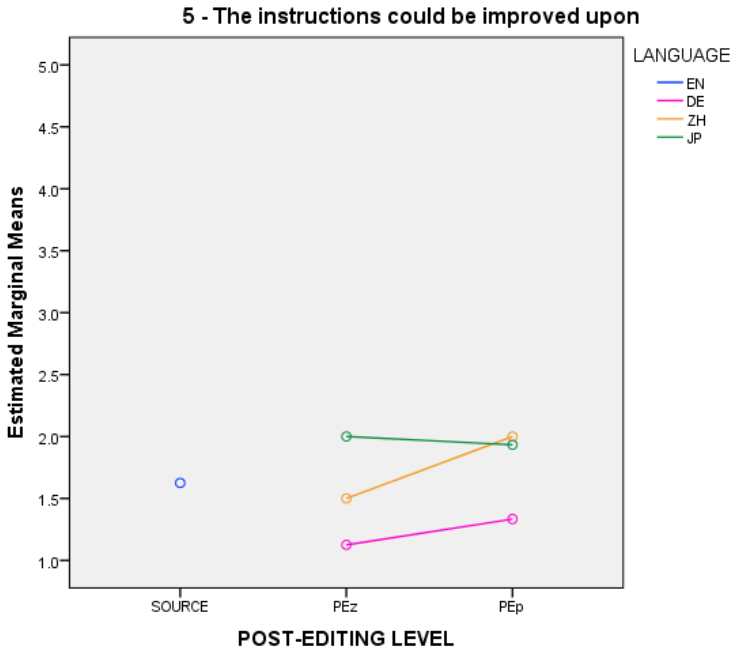


Figure 6:28 - Statement 5 - Source

Figure 6:29 illustrates the estimated marginal means for each translated language and their post-editing level compared to the EN_Source for statement 6.

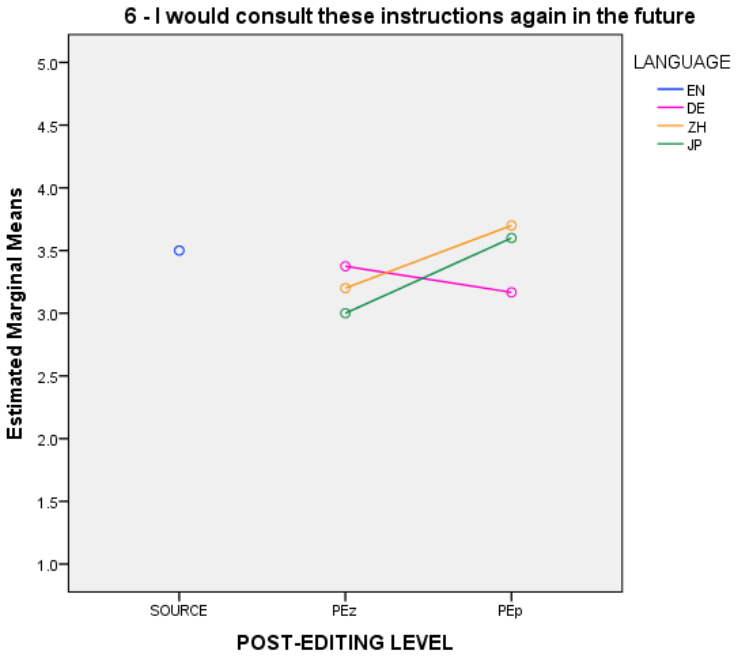


Figure 6:29 - Statement 6 - Source

A higher rating for statement 6 can be observed for EN_Source when comparing with the both German groups, ZH_PEz and JP_PEz groups as well. Interestingly, the Simplified Chinese and Japanese PEp groups show higher rating when compared to the EN_Source. However, none of these results were statistically significant (p>.10).

Figure 6:30 illustrates the estimated marginal means for each translated language and their post-editing level compared to the EN_Source for statement 7.

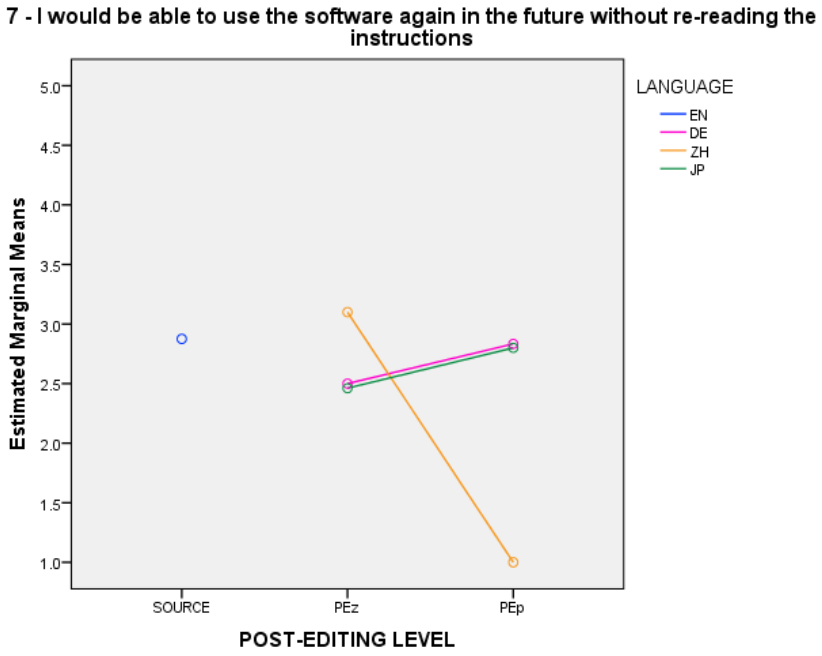


Figure 6:30 - Statement 7 - Source

The EN_Source (M=2.87, SE=.38) presents a higher rating when compared to the DE_PEz, DE_PEp, JP_PEz, JP_PEp and ZH_PEp groups which indicates that the EN_Source group considers that they would be more able to reuse the software without the instructions. This result was only statistically significant when compared to the ZH_PEp (M=1.00, SE=.34) group. The ZH_PEz however shows a slightly higher rating for statement 7 when compared to the EN_Source, but no statistically significant difference between the two groups were found (p >.10).

Figure 6:31 illustrates the estimated marginal means for each translated language and their post-editing level compared to the EN_Source for statement 9.

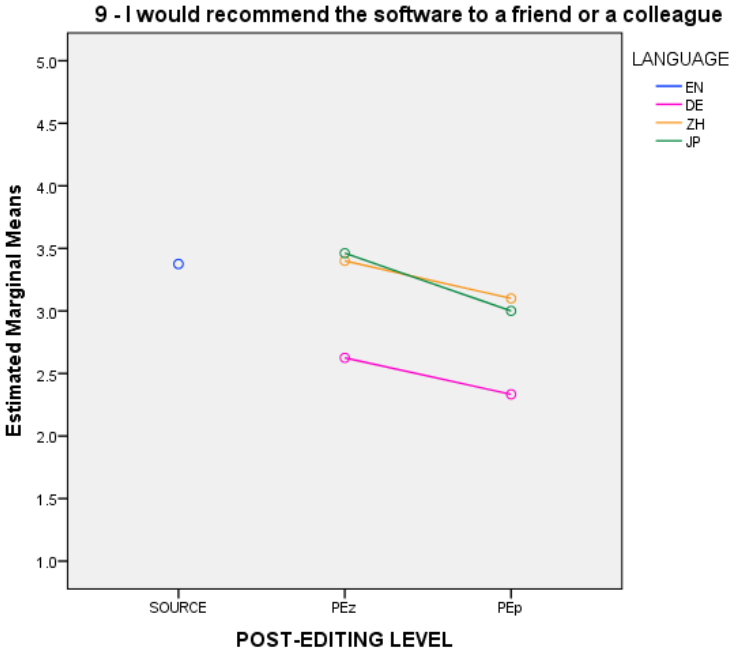


Figure 6:31 - Statement 9 - Source

The EN_Source (M=3.37, SE=.35) presents a higher rating when compared to the DE_PEz, DE_PEp, JP_PEp and ZH_PEp groups which indicates that the EN_Source group were more likely to recommend the product. This result was only statistically significant when compared to the DE_PEp (M=2.33., SE=.40) group. The JP_PEz however shows a slightly higher rating for statement 9 when compared to the EN_Source, but no statistically significant difference between the two groups was found (p >.10).

6.2.2 Moderators' ratings

6.2.2.1 MT Instructions

A two-way ANOVA with repeated measures was conducted in order to compare whether the factors Language and PE_LEVEL have an effect on the statements in the satisfaction of the moderators. As mentioned previously, the moderators were asked to assess the translated instructions regarding fluency, adequacy, grammar, style (see Section 6.1.2) 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.

LANGUAGE: The factor Language was found to have a statistically significant difference on satisfaction, where (F (2, 12) = 3.91, p<.05). This means that when the factor language is considered without distinctions between PE_LEVELs, there is a statistically significant differences across the three translated languages DE (M=1.7, SE=.11), ZH (M=1.4, SE=.11), JP (M=1.9, SE=.11).

POST-EDITING LEVEL: The factor PE_LEVEL was found to have a very statistically significant difference on satisfaction, where (F (1, 12) = 40.90, p<.001). This means that when the factor PE_LEVEL is considered without distinctions between languages, there is a statistically significant difference across the two post-editing levels PEz (M=1.29, SE=.09), and PEp (M=2.14, SE=.09).

INTERACTION: The interaction Language*PE_LEVEL was also found to have a statistically significant effect on satisfaction, where (F (2, 12) = 3.53, p<.05). This means that the factor language combined with the factor PE_LEVEL have a joint effect on satisfaction.

Table 6:9 shows the mean and standard deviation for each language and their respective post-editing levels.

Instructions Type		Mean	Std. Deviation
DE	PEz	1.11	0.10
	PEp	2.44	0.10
ZH	PEz	1.11	0.10
	PEp	1.83	0.50
JP	PEz	1.67	0.34
	PEp	2.16	0.29

Table 6:9 - Mean and Standard Deviation for Satisfaction - Moderators' rating – MT Instructions

A pairwise comparison found that there was a statistically significant difference between the Simplified Chinese language (no distinction between PE_LEVELs) when compared to the German language at the $p<.10$ level and Japanese language at the $p<.05$ level, where the Chinese language presented lower ratings for satisfactions. There was no statistically significant difference for the German language when compared to the Japanese language. Regarding the PE_LEVELs, there was also a highly statistically significant difference between PEz and PEp (no distinction between languages) at the $p<.001$ level, where the PEz presented lower ratings for satisfaction. Figure 6:32 illustrates the estimated marginal means for each translated language and their post-editing levels.

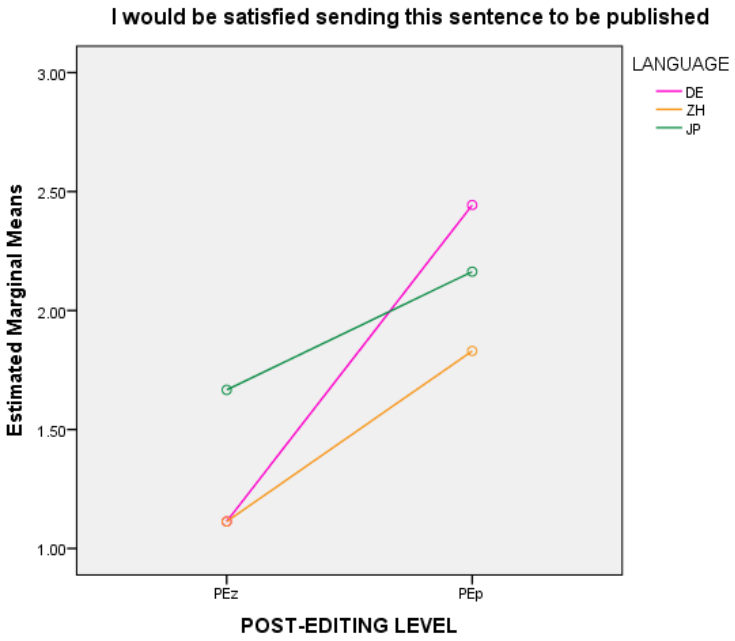


Figure 6:32 – Satisfaction – Moderators' ratings – MT Instructions

A higher rating for satisfaction can be observed for DE_PEp group when compared to ZH_PEp and JP_PEp groups, which indicates that the lightly post-edited version of the German language was more satisfactory for the moderators than the post-edited versions of Simplified Chinese and Japanese. However, this results was only statistically significant for the comparison DE_PEp against ZH_PEp at the $p<.05$ level. When looking at the PEz groups, a higher rating can be observed for the JP_PEz group when compared to the DE_PEz and ZH_PEz. This result was statistically significant at the p <.05 level when comparing JP_PEz against ZH_PEz and DE_PEz which means that among the raw machine translated instruction, the Japanese

language scored higher for satisfaction. Finally, when comparing the PEp instructions against their own PEz version, the DE_PEp, ZH_PEp and JP_PEp show higher ratings for satisfaction which indicates that the moderators were more satisfied in sending for publication the lightly post-edited version of the instructions. These results were very statistically significant for the DE_PEp at the p <.001 level, for the ZH_PEp and JP_PEp instructions at the p <.05 level.

6.2.2.2 HT Instructions

As stated previously, for the usability experiment, two sets of instructions were human translated and incorporated into the MT instructions as two control tasks (tasks 4 and 8). For the quality experiment, these two HT set of instructions were also displayed for moderators to rank in terms of satisfaction and were also incorporated into the MT instructions they were ranking. The moderators did not know what set of instructions were the HT instructions.

A one-way ANOVA was conducted in order to compare the effect of Instruction type (DE_HT, ZH_HT and JP_HT) on satisfaction, and found a statistically significant difference among the language, where (F (2, 15) = 6.40, p<.05).

Table 6:10 shows the mean and standard deviation, while Figure 6:33 illustrates the estimated marginal means for each language and Instruction type.

Instruction Type		Mean	Std. Deviation
DE	HT	2.42	0.20
ZH	HT	1.67	0.52
JP	HT	2.33	0.41

Table 6:10 - Mean and Standard Deviation for Satisfaction - Moderators' rating – HT Instructions

The highest rank for satisfaction can be observed for the German language, closely followed by the Japanese language. A statistically significant difference was found for the Simplified Chinese language when compared to the German language, at the p<.005 and Japanese language, at the p<.05 level, which indicates that Simplified Chinese moderators who assessed the two sets of human translated instructions were less satisfied sending the sentence to be published when compared to the German and Japanese moderators.

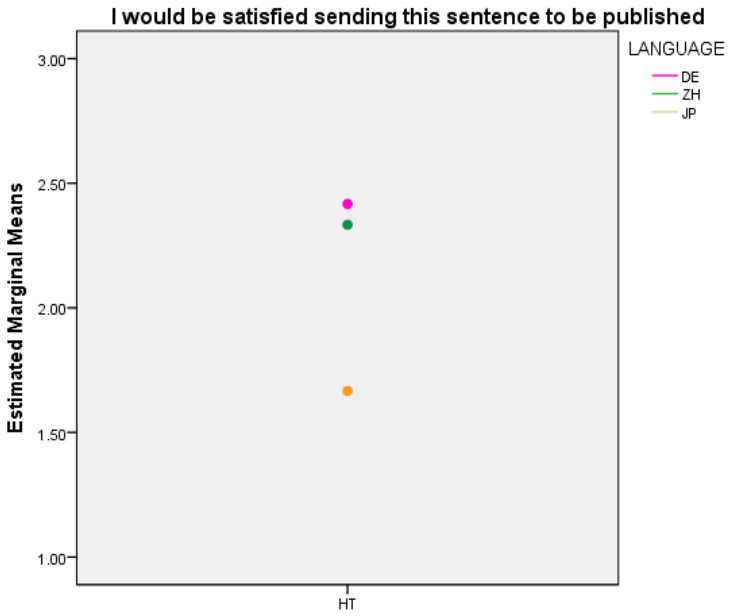


Figure 6:33 - Satisfaction - Moderators' rating - HT Instructions

6.2.2.3 MT Instructions vs HT Instructions

In order to identify whether there are differences between the ratings for the HT instructions and ratings for the MT Instructions (PEz and PEp), a two-way ANOVA was conducted. Table 6:11 shows the mean and standard deviation, while Figure 6:34 illustrates the estimated marginal means for each language and Instruction type.

Instruction Type		Mean	Std. Deviation
DE	PEz	1.11	0.10
	PEp	2.44	0.10
	HT	2.42	0.20
ZH	PEz	1.11	0.10
	PEp	1.83	0.50
	HT	1.67	0.52
JP	PEz	1.67	0.34
	PEp	2.16	0.29
	HT	2.33	0.41

Table 6:11 - Mean and Standard Deviation for Satisfaction - Moderators' rating

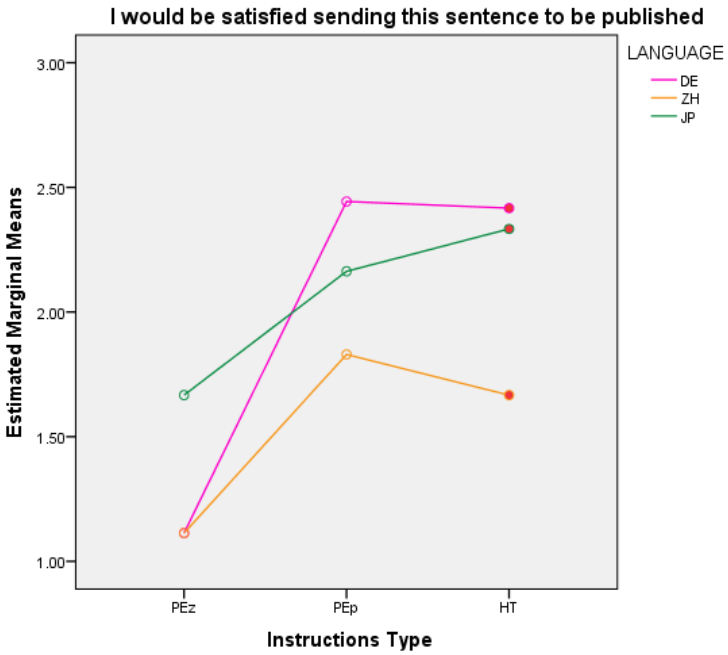


Figure 6:34 - Satisfaction - Moderators' rating - MT vs HT Instructions

When comparing the HT Instructions with MT Instructions for the German language, a slightly higher rating for the PEp Instructions can be observed, but no statistically significant differences between the two instruction types. When comparing the PEz and HT instructions for German, a very statistically significant difference was found at the $p<.001$ level. The results for German mean that moderators were less satisfied with the PEz instructions but comparatively satisfied with the PEp and HT Instructions.

When looking at the Japanese language, a higher ranking can be observed for the HT instructions when compared to the PEp, but no statistically significant differences the two instruction types were found. When comparing the PEz and HT instructions, a strong statistically significant difference was found at the $p<.05$ level. These results for Japanese mean that moderators were less satisfied with the PEz instructions but comparatively satisfied with the PEp and HT instructions.

Finally, for the Simplified Chinese language, a higher rating for the PEp Instructions can be observed when compared to the HT, but no statistically significant differences between the two instruction types were found. When comparing the PEz and HT instructions, a strong statistically significant difference was found at the $p<.05$ level. The results for Chinese follow the previous results for German and Japanese,

where moderators were less satisfied with the PEz instructions but comparatively satisfied with the PEp and HT Instructions.

6.2.3Web Survey

As described in Chapter 4 (Section 4.2.2.2.5), the web survey of satisfaction consisted of one single question: "Was this information helpful?" which could be answered simply with YES or NO. The scores illustrate the percentage of "YES" answered by the end users on the industry partner's webpage. Unfortunately, the rating numbers for each language and post-editing levels consist of sensitive information for the industry partner, and therefore, the absolute average will not be reported, but instead, a DELTA score is reported. The DELTA score is a subtraction of the score of one group from the other group, so the difference between two groups is calculated (e.g. JP_PEp score (% of yes) *minus* JP_PEz score (% of yes)) Therefore, in this section the DELTA scores for PEp vs PEz, HT vs PEp, HT vs PEz and EN vs HT are reported. The choice of comparing the EN with HT came from the understanding that a straight comparison with experimental PEz and PEp would not result in truthful results.

PEp vs PEz

The DELTA percentage of YES for the PEp and PEz instructions are displayed in Table 6:12. Note that for the German and Japanese languages, the PEp instructions have higher ratings when compared to the PEz instructions (PEp>PEz), but for the Simplified Chinese language, the PEz instructions have a higher rating when compared to the PEp (PEp<PEz). When looking at the DELTA, the difference for German is 1%, which means that the PEp instruction ratings were 1% higher than the PEz. For Japanese, that difference is slightly higher, at 2.38%. For the Simplified Chinese language, the PEz instruction ratings were 4.31% higher than the PEp.

Language	DE	ZH	JP
Average rating	PEp>PEz	PEp<PEz	PEp>PEz
DELTA (PEp - PEz)	1%	-4.31%	2.38%

Table 6:12 - DELTA scores Web Survey Satisfaction - PEp *vs* PEz

HT vs PEp

The DELTA percentage of YES for the HT and PEp instructions are displayed in Table 6:13. For all the languages, the HT instructions have higher ratings when compared to the PEp instructions (HT>PEp). When looking at the percentages, the difference for German shows that the HT instruction ratings were 3.45% higher than the PEp. For Japanese, that difference is slightly higher, at 3.47%. The Simplified Chinese language shows the highest difference, with the HT instruction ratings 5.81% higher than those of PEp.

Language	DE	ZH	JP
Average rating	HT>PEp	HT>PEp	HT>PEp
DELTA (HT - PEp)	3.45%	5.81%	3.47%

Table 6:13 - DELTA scores Web Survey Satisfaction - HT *vs* PEp

HT vs PEz

The DELTA percentage of YES for the HT and PEz instructions are displayed in Table 6:14. Similarly to the previous HT comparison, all the languages have higher ratings for the HT when compared to the PEz instructions (HT>PEz). When looking at the percentages, the difference for German is higher than the previous HT vs PEp comparison, since HT is 4.38% higher. For Japanese, that difference is even higher when compared to the previous HT vs PEp, since the HT instructions scored 5.85% higher than the PEz. For the Simplified Chinese language, an opposite trend is observed from the previous comparison HT vs PEp, as the difference in the ratings between HT vs PEz dropped to 1.50%.

Language	DE	ZH	JP
Average rating	HT>PEz	HT>PEz	HT>PEz
DELTA (HT - PEz)	4.38%	1.50%	5.85%

Table 6:14 - DELTA scores Web Survey Satisfaction - HT *vs* PEz

EN vs HT

The DELTA percentage of YES for the EN and HT instructions are displayed in Table 6:15. For the Chinese and Japanese languages, the HT instructions have higher ratings when compared to the EN source instructions (EN<HT), but for German language, the EN instructions have a higher rating when compared to the HT (EN>HT). When looking at the percentages, the English instructions shows 15.66% higher rating

when compared to the German HT instructions. For the Japanese language, the HT instructions show 4.79% higher ratings when compared to the EN source instructions. For Chinese, that difference is even higher, where the HT instructions are scored 14.81% higher than the English source.

Language	DE	ZH	JP
Average rating	EN>HT	EN<HT	EN<HT
DELTA (EN - HT)	15.66%	-14.81%	-4.79%

Table 6:15 - DELTA scores Web Survey Satisfaction - EN *vs* HT

Chapter 7 – Discussion

This Chapter discusses the findings reported in previous chapters (Chapter 5 and Chapter 6) regarding the usability, satisfaction and quality experiments. It starts with a discussion of the usability (effectiveness and efficiency) and cognitive data results (Section 7.1), followed by the satisfaction results (post-task questionnaire, web survey and moderators' ratings (TQA)) in Section 7.2, and ending with a discussion of the quality results in Section 7.3. The research questions are revisited in order to find whether they are answered by the findings presented.

7.1 Usability

Usability was measured via three elements: effectiveness, efficiency (goal completion/time) and satisfaction. Cognitive data was also gathered as an aid to understand the cognitive effort required to perform the eight tasks. Therefore, the discussion starts with the effectiveness and efficiency (task time, goal completion/task time) followed by the cognitive effort. The aim of the usability experiment was to answer the following research questions:

> ***RQ1****: Does Post-editing level have an effect on usability?*
>
> ***RQ4****: How do different target languages compare in terms of usability for both PEp and PEz content?*
>
> ***RQ7****: How does usability of Source Content compare with usability of the translated content (PEp and PEz)?*

Table 7:1 shows the summary of results for the usability measures. The tables should be read as mathematical symbols, i.e. the first row of the MT Instructions column shows PEp>PEz which means light post-editing groups showed greater effectiveness scores than their raw machine translation groups. The second row of the MT Instructions column shows PEp≠PEp, which means that statistically there was a difference among the PEp groups, where the German group presented statistically significant greater scores for effectiveness against Japanese and Simplified Chinese (DE > JP, ZH). The third row of the MT instructions shows PEz=PEz, meaning that there was

no statistically significant differences among the PEz groups for effectiveness. In the HT Instructions column, the first row shows PEp=PEz, which means that there was no statistically significant difference between the group of participants who used the HT instructions embedded in the PEp instructions and the group of participants who used the HT instruction embedded in the PEz instructions.

Measures	MT Instructions	HT Instructions
Effectiveness	PEp > PEz	PEp = PEz
	PEp ≠ PEp (DE > JP, ZH)	PEp = PEp
	PEz = PEz	PEz = PEz
	Source = PEp	Source = PEp
	Source = PEz	Source = PEz
Task Time	PEp = PEz (DE, JP); PEp < PEz (ZH)	PEp = PEz
	PEp = PEp	PEp = PEp
	PEz ≠ PEz (JP <ZH)	PEz = PEz
	Source = PEp	Source = PEp
	Source = PEz	Source = PEz
Efficiency	PEp = PEz (JP); PEp > PEz (DE, ZH)	PEp = PEz
	PEp = PEp	PEp ≠ PEp (JP> ZH)
	PEz = PEz	PEz ≠ PEz (JP >DE,ZH)
	Source = PEp	Source = PEp (DE, JP); Source > PEp (ZH)
	Source > PEz	Source = PEz (DE, JP); Source > PEz (ZH)

Table 7:1 - Summary of Results for Usability

The results for the MT Instructions show that the PEp groups are more effective, more efficient and faster than the PEz groups for the majority of cases, with the exception of task time for German and Japanese, and efficiency for Japanese where the PEp groups did not differ statistically from the PEz groups, but still show higher means. These results confirm that the factor post-editing level does have an effect on usability for all languages, where the implementation of light post-editing increased the effectiveness and efficiency of the instructions. For the HT instructions, when looking at the raw data, slightly higher means can be observed for the PEp group when using the HT instructions against the PEz groups. This could be due to the fact that, as the PEp group is more effective and efficient throughout the tasks, they may feel more confident performing the tasks. However, none of the PEp groups statistically differ from their PEz groups, for any language, for any of the measures with HT Instructions. This result confirms that when provided with the same instructions, no statistically significant differences are found between PEp and PEz groups of any languages. Statistically significant differences are only seen when groups use raw machine

translated or lightly post-edited instructions (MT Instructions), confirming again that post-editing had an effect on usability and thus, answering the RQ1.

Regarding how different languages compare in terms of usability (RQ4), it is interesting that the Simplified Chinese language always shows differences between PEp and PEz groups for the MT instructions. Also, the PEz group of the Simplified Chinese language seems to generally score lower, closely followed by DE_PEz group. The Japanese group shows fewer differences between its PEp and PEz groups, often scoring in the middle and performs highest when using the HT instructions compared to when they use the MT instructions. These results indicate that the implementation of light post-editing seem to affect more the Simplified Chinese and German languages regarding usability, than the Japanese language.

When considering RQ7, it is possible to say that the EN_Source group presents the same levels of goal completion, task time and efficiency as the PEp groups as the EN_Source does not statistically differ from any languages. These results confirm that the level of usability of the source can be directly compared to that of the PEp groups of all languages. When looking at the HT instructions, the EN_Source is also comparable to the PEp groups for all measures, apart from the Simplified Chinese language for efficiency, which is a bit surprising, as one would expect that by using the high quality instructions, the Simplified Chinese language would perform similarly to all the other languages, including the EN_Source. This result can be explained by looking at the results of the quality experiment performed by the moderators (TQA) where the HT instructions of the Simplified Chinese language scored lowest for all the categories (except for 'spelling' where Simplified Chinese is scored second). Moreover, the Simplified Chinese HT instructions were scored the same as the PEz instructions for spelling and sentence structure and, although not statistically significant, the Simplified Chinese HT instructions showed *lower* scores for the adequacy and terminology categories when compared to the PEp instructions. On the basis of these results, it is possible to affirm that the HT instructions of the Simplified Chinese did not show the quality expected, which influenced the efficiency measure. When comparing the EN_Source to the PEz groups, it can be affirmed that statistically the EN_Source group presents the same levels of goal completion and task time when looking at both MT and HT instructions, even though the EN_Source group shows higher means. However, EN_Source shows higher efficiency when compared to all PEz groups of all languages

for the MT instructions. This is due to the fact that efficiency takes into consideration successful tasks divided by task time and, as mentioned above, the EN_Source was slightly faster and slightly more effective (even if not statistically significant) and therefore, the efficiency scores were higher. For the HT instructions, EN_Source was once again only different from the Simplified Chinese language. These results also confirm the results discussed above, in which the HT instructions of Simplified Chinese were not of expected quality. In conclusion, the RQ7 can be answered as: the results show that the levels of usability for the source instructions can be directly compared to that of the lightly post-edited instructions, but not directly compared to the PEz groups, since EN_Source shows higher efficiency scores. Thus, even light PE can bring raw MT output to a level of usability comparable to a good quality source text.

Cognitive Data

The results for the cognitive data regarding RQ1 are presented in Table 7:2.

Measures	AOI	MT Instructions	HT Instructions
Fixation Duration	INSTR	PEp = PEz (DE,JP); PEp < PEz (ZH)	PEp = PEz
	UI	PEp = PEz	PEp = PEz
Fixation Count	INSTR	PEp = PEz	PEp = PEz
	UI	PEp = PEz	PEp = PEz
Visit Count	INSTR	PEp = PEz (DE, JP); PEp < PEz (ZH)	PEp = PEz
	UI	PEp = PEz (DE, JP); PEp < PEz (ZH)	PEp = PEz

Table 7:2 - Cognitive Effort - PE_Level

Statistically, only the Simplified Chinese language shows more cognitive effort required from the PEz group when compared to the PEp group for both AOIs. However, all PEz groups show higher means for fixation duration, fixation count and visits when compared to their respective PEp groups for the AOI INSTR (not statistically significant). This result correlates with the results for usability and helps to confirm that post-editing level has an effect on cognitive effort when reading the instructions. For the AOI UI, although only statistically significant for the Simplified Chinese language, in the majority of the cases the PEz groups present higher means for fixation duration, count and visits when compared to the PEp groups. The exception is the Japanese language whose PEp group presents higher means for fixations (duration and count) when compared to the PEz group and the German language whose PEp group presents slightly higher mean for visit counts than the PEz group.

The issue with German and Japanese might have to do with the fact that the terminology in the UI did not seem to be intuitive: the Japanese instructions had some terminology (font names) in Japanese characters while the UI presented the roman characters for the same term. For the German language, the UI terminology is sometimes shorter than the instructions because of lack of space (a typical issue in localisation): e.g., in task 5, the command "Wählen Sie Grüne Füllung mit dunkelgrünem Text" was presented in the instructions, while in the UI only "Grüne Füllung" could be seen. Some participants clicked 'edit' and tried to get the dark green text – as it was not intuitive that the "Grüne Füllung" option would already change the text to dark green. One would then expect that both PEz and PEp groups would have the same problems with this terminology, however, since the PEp groups have more completed tasks, it is possible that those groups spent more time in the UI trying to look for the right terminology.

For the HT instruction, none of the PEp groups statistically differ from their PEz groups, for any language, for any of the measures, which is an expected result since the instructions seen by PEp and PEz groups were the same. However, one may observe a mixture of results (even if not statistically significant) where the PEp groups show higher fixation duration time and a higher number of visits than the PEz groups in the AOI UI, but the PEz groups show higher fixation count. It is, however, not clear why differences happened with the HT Instructions, but this may be due to the fact that the sample size was not big enough for showing more statistically significant patterns. In conclusion, the cognitive data regarding post-editing level, although presenting mixed results, helps to answer the RQ1 where the post-editing has a statistically significant effect for the Simplified Chinese language, and a trend for the German and Japanese languages.

Table 7:3 shows the summary of statistically significant results for the cognitive data regarding the RQ4.

Measures	AOI	MT Instructions	HT Instructions
Fixation Duration	INSTR	PEp = PEp	PEp ≠ PEp (DE > JP)
		PEz = PEz	PEz ≠ PEz (ZH > JP)
	UI	PEp = PEp	PEp = PEp
		PEz = PEz	PEz = PEz
Fixation Count	INSTR	PEp ≠ PEp (DE > ZH)	PEp ≠ PEp (DE > ZH, JP)
		PEz = PEz	PEz = PEz
	UI	PEp = PEp	PEp = PEp
		PEz = PEz	PEz = PEz
Visit Count	INSTR	PEp = PEp	PEp = PEp
		PEz ≠ PEz (ZH> DE)	PEz ≠ PEz (ZH> DE)
	UI	PEp = PEp	PEp = PEp
		PEz ≠ PEz (ZH> DE)	PEz ≠ PEz (ZH> DE)

Table 7:3 - Cognitive Effort - Language

The statistically significant results show that there are not many differences among the languages when using the MT Instructions, only for the German PEp group who show more fixation counts in the AOI Instructions than the Simplified Chinese PEp groups; and for the Simplified Chinese PEz group who shows more visits in the AOI INSTR and UI than the German PEz group. The only distinct pattern observed was that of the PEz group for Simplified Chinese that shows higher fixations and visits the majority of times when compared to the other PEz groups. This result correlates well with the usability findings, where the PEz group of Simplified Chinese seemed to score lower most of the time as well.

The results for the HT instructions are more mixed, where the PEp group of the German language shows statistically higher fixation duration time and more fixation counts in the AOI INSTR, and the PEz group of the Simplified Chinese shows higher fixation duration time and visit counts in the AOI INSTR and more visits in the AOI UI when compared to the others PEz groups. By looking at the data, the only distinctive pattern that can be observed is that again, the PEz group of the Simplified Chinese language shows longer fixation duration times and higher visit counts when compared to the other languages. The findings for the HT instructions and Chinese might be also explained by the fact that the Chinese HT instructions did not show the highest quality in the TQA evaluation. Regarding the PEp group of the German showing more fixation counts in the AOI INSTR, it could also be connected to the fact that they had more tasks successfully completed and, therefore, relied more on those instructions. Based on these results, it is possible to affirm that the PEz group of the Simplified Chinese

language showed more cognitive effort when compared to the other groups, thus confirming that there is a difference among languages (RQ4). It is important to highlight once again that the results for the HT Instructions are not very clear and it may be due to the sample size.

The statistically significant results for the cognitive data regarding the source content (RQ7) are presented in Table 7:4.

Measures	AOI	MT Instructions	HT Instructions
Fixation Duration	INSTR	Source = PEp (ZH,JP); Source<PEp (DE)	Source = PEp (JP); Source<PEp
		Source < PEz	Source = PEz (JP); Source<PEz
	UI	Source = PEp (ZH,DE); Source<PEp (JP)	Source = PEp (JP,DE); Source<PEp
		Source = PEz (DE,JP); Source<PEz (ZH)	Source = PEz (JP,DE); Source<PEz
Fixation Count	INSTR	Source = PEp (JP,ZH); Source<PEp (DE)	Source = PEp (JP,ZH); Source<PEp
		Source = PEz (ZH); Source<PEz (DE,JP)	Source = PEz (JP,ZH); Source<PEz
	UI	Source = PEp	Source = PEp (DE,JP); Source<PEp
		Source = PEz (DE,JP); Source< PEz (ZH)	Source = PEz
Visit Count	INSTR	Source = PEp	Source = PEp
		Source = PEz (DE); Source<PEz (JP,ZH)	Source = PEz (DE); Source<PEz
	UI	Source = PEp	Source = PEp
		Source = PEz (DE,JP); Source<PEz (ZH)	Source = PEz (DE,JP); Source<PEz

Table 7:4 - Cognitive Effort - Source

Statistically, the EN_Source most of the time presents the same amount cognitive effort as the PEp instructions for both MT and HT Instructions, for both AOIs, in keeping with usability findings. The few differences are against the German PEp group for fixations (duration and count) in the AOI INSTR, and against the Japanese group for fixation duration in the AOI UI. Against the PEz groups, the source mostly shows differences against the Simplified Chinese, but some differences against German and Japanese PEz groups can also be observed. The differences for the HT instructions mostly lie against both groups (PEp and PEz) of the Simplified Chinese language (both AOIs), and the German groups (PEp and PEz) for the AOI INSTR. These results help to answer RQ7: in general, the EN_Source shows a lower level of cognitive effort required, but it is quite similar to the PEp instructions, whereas there are greater differences compared with the PEz instructions.

In conclusion, it has been shown that post-editing level has an effect on usability (RQ1), where the lightly post-edited instructions show the higher levels of effectiveness and efficiency when compared to the raw machine translated instructions. The cognitive data, although presenting somewhat mixed results, correlates with the usability result, where more cognitive effort could be observed in

general for the PEz groups. Regarding how different languages compare in terms of usability (RQ4), results have shown that the simplified Chinese and the German languages are more affected by the implementation of light post-editing than the Japanese language. The cognitive data showed patterns that the PEz instructions of the Simplified Chinese requires more cognitive effort in the majority of case, corroborating for the finding that by implementing post-editing for Chinese it increases the usability and decreases the cognitive effort. Finally, when analysing how usable the source content is compared to the translated content (RQ7), it has been shown that the level of usability of the source can be directly compared to the PEp groups, but the source usability is higher when compared to the PEz groups. The cognitive data showed patterns where the EN_Source has a lower level of cognitive effort required, which can be closely compared to the PEp instructions, but not to the PEz instructions, which required more cognitive effort.

7.2 Satisfaction

The aim of the satisfaction experiments was to answer the following research questions:

> ***RQ2****: Does Post-editing Level have an effect on satisfaction?*
>
> ***RQ5****: How do different target languages compare in terms of satisfaction for both PEp and PEz content?*
>
> ***RQ8****: How does satisfaction with Source Content compare with satisfaction with translated content (PEp and PEz)?*

Post-task Questionnaire

Table 7:5 shows the summary of statistically significant results.

Statements[37]	
1. The instructions were usable.	PEp = PEz (JP, DE); PEp > PEz (ZH)
	PEp = PEp
	PEz = PEz
	Source = PEp
	Source > PEz
2. The instructions were comprehensible.	PEp = PEz (JP, DE); PEp > PEz (ZH)
	PEp = PEp
	PEz = PEz
	Source = PEp
	Source > PEz
3. The instructions allowed me to complete all of the necessary tasks	PEp = PEz (JP); PEp > PEz (ZH, DE)
	PEp = PEp
	PEz = PEz
	Source = PEp
	Source = PEz
4. I was satisfied with the instructions provided.	PEp = PEz
	PEp = PEp
	PEz = PEz
	Source = PEp
	Source = PEz (JP, ZH); Source > PEz (DE)
5. The instructions could be improved upon.	PEp = PEz
	PEp = PEp
	PEz = PEz
	Source = PEp
	Source = PEz
6. I would consult these instructions again in the future	PEp = PEz
	PEp = PEp
	PEz = PEz
	Source = PEp
	Source = PEz
7. I would be able to use the software again in the future without re-reading the instructions.	PEp = PEz (JP, DE); PEp < PEz (ZH)
	PEp ≠ PEp (ZH < DE, JP)
	PEz ≠ PEz (ZH < DE, JP)
	Source = PEp (DE, JP); Source > PEp (ZH)
	Source = PEz
8. I would rather have seen the source (English) version of the instructions	PEp = PEz (JP); PEp > PEz (ZH, DE)
	PEp = PEp
	PEz ≠ PEz (JP < DE, ZH)
9. I would recommend the software to a friend or a colleague.	PEp = PEz
	PEp = PEp
	PEz ≠ PEz (JP > DE)
	Source = PEp (JP, ZH); Source > PEp (DE)
	Source = PEz

Table 7:5 – Summary of Results for the Post-Task Satisfaction Questionnaire

[37] Lower rating for statement 5 and 8 indicate more agreement towards them.

The results for the post-task satisfaction questionnaire show that for the majority of the statements the PEp instructions were scored the same as the PEz instructions, with the differences between PEp and PEz mostly observed for the Simplified Chinese language (statements 1, 2, 3, and 8) and some for the German language (statements 3 and 8). The result for German regarding statement 3 are interesting since the German PEz group was able to complete more tasks when compared to the other PEz groups, yet the instructions were not considered very usable by the participants who used the DE_PEz instructions. Japanese is the only language that did not present statistically significant differences between the PEp and PEz groups for any of the statements. The PEp groups of all languages score higher for statements 1, 2, 3, 4 and 8 (even if no statistically significant differences are found) when compared to their respective PEz group. These results help to answer RQ2 by confirming that there was a difference in the perceived satisfaction between the PEp and PEz groups, especially for the Simplified Chinese language, which in turn, correlates well with the usability experiments where the Simplified Chinese language always shows differences between PEp and PEz groups for the MT instructions, and has its PEz group generally scored lower than the other PEz groups; this also correlates with the cognitive data where the Simplified Chinese language is the only language showing statistically significant difference in cognitive effort between PEp and PEz groups (Table 7:2).

Regarding how languages differ in terms of satisfaction (RQ5), it is interesting that the Simplified Chinese language shows greater differences between PEp and PEz groups but it is the DE_PEz group that seems to be the least satisfied with the instructions even if PEz instructions were scored the same as the PEp instructions, closely followed by the ZH_PEz group. Very similarly to the usability results, the Japanese language shows fewer differences between its PEp and PEz groups, often being scored somewhere in the middle (JP_PEp scoring higher, closely followed by the JP_PEz). Additionally, the Japanese language groups seemed to be less interested in working with the source (statement 8). One point to be mentioned is the results for statement 5 (the instructions could be improved upon) which was rated low for all languages for all the groups including the source – where the highest mean is 2.00 (in a 1-5 Likert scale). This is very interesting especially because the participants were also taking into consideration the two human translated instructions they used to perform the usability tasks, which showed high levels of effectiveness (goal completion).

When considering how the source compared to the other groups (RQ8), it can be observed that the source mostly does not differ from the PEp groups, a result that aligns well with the previously reported results. Significant differences are only observed against the Simplified Chinese (statement 7) and German (statement 9) groups. When looking at the raw data, it is interesting that the EN_Source scored highest only for statements 1, 2 and 4 (of all PEp and PEz groups), indicating that satisfaction with the source content was not absolutely higher than all groups. More interesting is that for statement 5, the source is in the middle, indicating that the participants that used the source considered it to need more improvement than the participants who used the Japanese instructions (PEp and PEz) and the Simplified Chinese PEp instructions. This is very surprising as the EN_Source group presented a high level of usability and less cognitive effort when compared to the PEz groups and therefore, one would expect participants to respond that it did not need much improvement and to be more satisfied. When the source is compared to the PEz groups there are more statistically significant differences, where source is scored higher than PEz groups for statements 1, 2, 3, 4, 6, thus indicating that the satisfaction level for the source is higher than the satisfaction level for the PEz instructions.

Another interesting point concerns the pre-task survey results. As seen in Section 5.1, Chapter 5, the German participants present the highest levels of proficiency between C2-Proficiency level to C1-Advanced level for a greater part of the participants, whereas the Simplified Chinese and Japanese languages have the majority of participants between C1-advanced to B1-Intermediate levels. These results correlate with the post-task satisfaction questionnaire result, where the DE_PEz group were the least satisfied with the instructions among all the other groups (even if no statistically significant differences were found). This may be due to the fact that, because all the German participants have a good command of the English language they would prefer to read the instructions in their second language when given a bad quality translation. These results from the pre-task survey also correlate with the Japanese PEp and PEz groups answers for statement 8 (I would rather have seen the source instructions) where they displayed less agreement with the statement and were the only language group that did not present any statistically significant differences for all the statements of the post-task questionnaire. This may be because, with the average proficiency

between B2 upper intermediate and B1 intermediate, the Japanese participants were still more likely to use the translations than the English source. The results also correlate with the Simplified Chinese PEp group, which was the one that rated least likely to use the English source, even though the PEz group was the second most interested in using the source. Here we can speculate that the PEp instructions were perceived as a better option than using the instructions in their second language.

Web Survey Satisfaction

The Web Survey Satisfaction "Was this information helpful?" attempted to gather an indication of satisfaction levels from the real-end user of those articles. Table 7:6 presents a summary of the results for the web survey.

MT Instructions	HT Instructions
PEp>PEz (DE, JP)	HT>PEp
PEp<PEz (ZH)	HT>PEz
	Source>HT (DE)
	Source<HT (ZH, JP)

Table 7:6 - Summary of Results for the Satisfaction Web Survey

The results show that the PEp articles were rated higher for German and Japanese languages when compared to the PEz articles, which correlates with the usability results. For the Chinese language, interestingly, the PEp articles scored lower when compared to the PEz articles. Regarding the HT articles, all languages showed higher ratings for the HT articles when compared to both PEp and PEz articles. The results for the HT articles are unsurprising since their quality is expected to be higher. These results correlate with usability and post-task questionnaire results (RQ2) for German and Japanese, but differ with regards to the Simplified Chinese language.

Regarding differences among languages (RQ5), although it is not possible to report the absolute results due to confidentiality requested by the industry partner, it can be affirmed that the Simplified Chinese language shows the highest scores when compared to the other languages, for both PEp and PEz articles, closely followed by the Japanese language. The German language shows the lowest scores for the PEp and PEz articles. Even though the absolute results could not be reported, it can be affirmed that when the HT articles are compared across languages, the German language also seems to result in lower ratings for the online articles. The results for the German PEz

group and Japanese language follow the results of the post-task questionnaire where the DE_PEz participants seemed least satisfied while the Japanese scored medium. Different from the post-task questionnaire are the results for the PEz articles for the Simplified Chinese language, which tended to score the lowest in previous results.

Regarding the EN_Source articles (RQ8), the results show that they were rated 15% higher when compared to the HT articles of the German language. Interestingly, for Simplified Chinese and Japanese, the results are the opposite, where the source articles were rated 14% lower when compared to the HT articles of the Simplified Chinese language and 4% lower when compared to the HT articles of the Japanese language. Once again, the Japanese language seems to have the closest scores to the source, whereas the German language seems to have the lowest scores.

It is important to highlight here that, due to the nature of survey question, one has to consider that perhaps what is being rated in this survey is much more the content (if that information was helpful or not) rather than the translation itself. For example, the results for the Simplified Chinese PEz articles are opposite to what the usability, cognitive data, and post-task questionnaire have shown and, therefore, it is important to consider that the end users were rating if the content helped with their question instead of the quality of the translation. What this also shows is that one question on usefulness may not accurately capture actual acceptability by end users. It is also speculated that the cultural differences (cultural usability) may play a role in ratings for the differences languages. According to Suojanen, Koskinen and Tuominen (2015, p.21), “our cultural background may well affect what we understand to be usable and why”. As pointed out by the industry partner, it is known that German users tend to rate lower than English users, for example. Moreover, it has to be considered that it is not only English native speakers who use the source articles but also users from all over the world with different native languages who prefer to use the English source. This is confirmed by the industry partner who pointed out that ratings for the English source articles derived from Germany are also lower when compared to ratings for the English source articles coming from countries with primarily English native speakers. Therefore, the web survey question may not fully cover the satisfaction of the end user regarding the translation quality, but can be used as an indication of it.

TQA Satisfaction

A question about satisfaction was also included in the TQA (reported in Section 5.4.2 of Chapter 4).

MT Instructions	HT Instruction
PEp > PEz	HT = PEp
PEp ≠ Pep (DE>ZH)	HT > PEz
PEz ≠ PEz (JP>DE, ZH)	HT ≠ HT (DE>ZH)

Table 7:7 - Summary of Results for the Moderators' Satisfaction

The results for the question "I would be satisfied sending this sentence to be published" confirmed that PEp instructions were statistically more satisfactory for the moderators than the PEz instructions for all the translated languages. Regarding the HT instructions, they were statistically more satisfactory for all languages when compared to the PEz instructions. The PEp instructions were as satisfactory as the HT; that is, there were no statistically significant differences between the two instruction types across all languages. Interestingly, the PEp instructions for German and Simplified Chinese show higher means than the HT instructions. These results confirm that the implementation of light post-editing increased the quality of the instructions, even to the level of HT instructions for the Chinese and German languages, correlating with previous assessments of usability and post-task satisfaction questionnaire where the German and Simplified Chinese language seem to be more affected by the implementation of post-editing.

Concerning differences among languages (RQ5), the Simplified Chinese language showed the lowest scores for satisfaction (even though in some cases the difference was not statistically significant) among the PEp instructions. The PEz instructions of Simplified Chinese share the position with the PEz instructions of the German language where both languages show the lowest mean. The German language shows the highest scores among the PEp and HT instructions, whereas the Japanese language shows the highest score among the PEz instructions. These results confirm previous results where the Simplified Chinese and German PEz groups show the lowest scores among all the other instructions, and the Japanese language showing a medium score.

7.3 Quality

The quality of the source was measured with the help of two tools (see Section 4.2.2.2), whereas the quality of the translated content was measured via a TQA (see section 4.3.2.2.5). The aim of the quality experiments was to answer the following research questions:

RQ3: *Does the quality evaluation of Post-editing levels PEz and PEp, performed by professional evaluators reflect the results from the empirical usability and satisfaction experiments?*

RQ6*: Does the quality evaluation of the translated languages performed by professional evaluators reflect the results from the empirical usability and satisfaction experiments for Language?*

RQ09*: Does the quality evaluation of the Source Content reflect the results from the empirical usability and satisfaction experiments for Source Content?*

Measures	MT Instructions	MT vs HT Instructions
Adequacy	PEp > PEz	HT = PEp
		HT > PEz
	PEp ≠ PEp (DE > JP, ZH)	HT = HT (JP= DE, ZH)
	PEz ≠ PEz (DE, JP> ZH)	HT ≠ HT (DE>ZH)
Fluency	PEp > PEz	HT = PEp
		HT > PEz
	PEp ≠ PEp (DE, JP> ZH)	HT = HT
	PEz ≠ PEz (ZH, JP> DE)	
Spelling	PEp = PEz	HT = PEp
		HT = PEz (DE, ZH); HT > PEz
	PEp = PEp	HT = HT
	PEz = PEz	
Sentence Structure	PEp> PEz	HT = PEp
		HT = PEz (ZH); HT > PEz (DE,
	PEp = PEp	HT = HT
	PEz ≠ PEz (ZH, JP> DE)	
Terminology	PEp = PEz (DE, JP); PEp> PEz (ZH)	HT = PEp
		HT = PEz (ZH); HT > PEz (DE,
	PEp = PEp	HT = HT (JP= DE)
	PEz = PEz	HT ≠ HT (JP, DE>ZH)
Country Standards	PEp = PEz (JP); PEp> PEz (DE, ZH)	HT = PEp
		HT = PEz
	PEp = PEp	HT = HT (DE= JP, ZH)
	PEz ≠ PEz (JP> ZH)	HT ≠ HT (JP>ZH)

Table 7:8 - Summary of Results for the TQA

With regards to the quality of PEp and PEz instructions (RQ3), the PEp instructions were statistically higher scored than the PEz instructions for adequacy, fluency and sentence structure. The PEp instructions show higher scores in all measures when compared to the PEz instructions, apart from for the Simplified Chinese language for spelling, which shows the same mean for PEp and PEz instructions. Results also show that, statistically there were no significant differences between PEp and HT instructions for any of the measures. Moreover, the PEp instructions show higher ratings (even if not statistically significant) than the HT instructions for adequacy, sentence structure and country standards (for German and Chinese languages), and fluency (for German and Japanese languages). Against the PEz instructions, the HT instructions score higher the majority of the time (even if not statistically significant) for all measures and for all languages, however, the PEz instruction did not statistically differ from HT for country standards (any languages), spelling (for German and Chinese languages) and terminology (Chinese language). These results indicate that PEz instructions show lower quality than the HT instructions in general, yet they can also show similar quality for some measures (country standards, spelling and terminology in this case). Furthermore, the results of the TQA reflect the usability and satisfaction results for the translated content in terms that the implementation of post-editing increased the usability of the translation as well as the satisfaction as perceived by users and moderators, and in the case of the quality experiment, to the point that PEp instructions did not differ from the assessment for the HT instructions and sometimes even scored higher. Nonetheless, the raw machine translation versions also showed good scores especially for terminology, country standards and spelling. We consider that these results are likely because the industry partner has their own MT systems, which are trained in 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 participants who used the PEz instructions were still able to complete a great number of tasks even though more time and more cognitive effort was required.

Regarding the differences between the languages (RQ6), the PEp instructions only show statistically significant differences for adequacy – where the German language shows higher scores than the Japanese and Chinese languages; and fluency measures – where the Simplified Chinese language has lower scores than the other

languages. The PEz instructions show more statistically significant differences among the groups where the German or the Simplified Chinese groups show lower scores. These findings comply with previous results of usability and satisfaction. Regarding the HT instructions, the Japanese language once again shows higher scores when compared to the other HT groups – which correlates with previous usability results in which Japanese participants perform better when using HT instructions when compared to PEp and PEz instructions, as well as when compared to the participants from German and Chinese who used the HT instructions – whereas the Chinese HT instructions show the lowest scores. At this point it is worth recalling that for the Simplified Chinese language, the PEp instructions were scored higher than the HT instructions (even when it was not statistically significant) for adequacy, sentence structure, terminology and country standards, which helps to explain the previous results for efficiency in which the participants who used the HT instructions in Chinese showed statistically significant lower efficiency scores than the participants who used the source.

Finally, regarding the RQ09 – whether the quality evaluation of the source reflects the results of usability and satisfaction – it is possible to say that the source content contains the terminology and sentence structure which meets the expectation of what an end user would require when searching for instructions describing the features of the spreadsheet software; contained few issues regarding its features and retrieved a high Flesch Reading Ease score, and therefore, the quality of the source content can be considered of good quality, and simple and easy to read. Based on this, when looking back at the usability and satisfaction results, it is possible to see that the source content showed a high level of usability (comparable to the PEp instructions) and that the level of satisfaction with the source (post-task questionnaire) was also high. Nonetheless, the source content was also considered to be in need of improvement (statement 5).

7.4 Conclusion

This chapter discussed the results of the findings reported in the previous chapters. The implications of these findings, limitations and future potential work will be discussed in the final chapter.

Chapter 8 – Conclusions

The aim of this research was to answer the overarching question:

> *RQ: What factors influence acceptability levels of a machine translated text for the end user?*

The results for usability, quality and satisfaction have demonstrated that the factor Post-Editing Level had a significant effect on acceptability, where the lightly post-edited versions presented higher levels of acceptability when compared to the raw machine translated versions. Moreover, the lightly post-edited version showed a similar – or sometimes even higher – level of acceptability as the source content. Nonetheless, the raw machine translation versions were still usable and participants who used those versions of the instructions were still able to perform tasks. This result is comparable to that of Doherty and O'Brien (2014) in which the raw machine translated versions were also deemed usable in real-world scenarios.

Another factor that was found to have influenced acceptability is Language. It has been shown that the German and Simplified Chinese languages had greater levels of acceptability regarding their lightly post-edited versions compared to the raw MT versions. The findings were less clear-cut for the Japanese language.

The Source Content showed a high level of acceptability for all the elements (usability, quality and satisfaction), and was closely comparable to the lightly post-edited versions but higher than the raw machine translated output.

In summary, this study has shown that the implementation of *light* post-editing directly and positively influenced acceptability for German and Simplified Chinese languages, more so than for the Japanese language and, moreover, the findings of this research show that different languages have different thresholds for 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 in order to move to a more dynamic quality model (TAUS, QTLaunchpad, QT21) are now taking into consideration different views

of translation quality (DQF and MQM models) including the view of end users as suggested by the broad definition of translation by Koby et al. (2014) and the UCT approach, and therefore, this study corroborates to these attempts by demonstrating that translation quality should empirically factor in end user perceptions and ability to use content.

8.1 Limitations

The relatively small number of participants recruited for the usability experiments in which cognitive data was collected is a limitation of this work. Although when compared to previous work this study presented larger sample sizes for the translated content, the relatively small sample size is not optimal for a robust statistical analysis. The issues with the sample size were due to the fact that: i) the limited number of native speaker participants available in the place of data collection; ii) the volunteer nature of the participation did not allow any payment for the participants, relying on participants' good will for helping; and iii) a great number of recordings had to be discarded due to low quality (lower than 80%) – this was due to the fact that some of the participants wearing glasses or lenses, or even natural unsuitability for eye tracker such as shape of eyes, long eyelashes, etc. These limitations are countered by the fact that the number of participants was still adequate for the usability experiments and satisfaction (post-task questionnaire) rating. Moreover, the fact that this study was conducted in collaboration with an industry partner, meant that it was able to a) gather a great number of online ratings for the web survey as well as b) have eighteen professional moderators rating the quality of the translated content and c) use a domain-specific MT engine and content, all of which add to the ecological validity of the research.

Another point to highlight is the fact that only one content type is used in the research and, even though the pilot experiment was in collaboration with a different industry partner – and therefore they profile their content a bit differently – the type of content used in both pilot and main experiments can be described as instructional content. The reasons for using just one type of content have to do with the issue of finding a second content type, which would allow for the implementation tasks for the

usability experiments, and at the same time a content type which the industry partner was allowed to publish in the raw machine translation version and lightly post-edited version.

The methodology applied in this research could have benefited from a randomisation of the tasks in terms of post-editing level so that each participant could perform the tasks using the lightly post-editing, raw machine translated, and human translated instructions. Unfortunately, the eye tracker used in this research did not allow for the tasks to be randomised without tasks being presented out of the order in which some needed to be presented (i.e. task 4 needed to be presented before task 8, as well as task 6 needed to be presented before task 7).

8.2 Contributions

Despite the few limitations presented above, this study has been successful in several points, including answering the main research question by applying an adapted model inspired by the UCT approach and usability for the translation quality evaluation of machine translation.

As discussed in Chapter 2, both academia and translation industry largely disregard the end user of translations, mostly focusing either on the pedagogy and theory of translation quality or on evaluation models heavily based on error typology, respectively. The field of machine translation also does not take into consideration the end user and the few studies which have attempted to understand how the user of those translations interact with it, generally focus on reading comprehension, in which either the user answers comprehension questions about the text (and so the number of correct questions are counted), or the user reads the texts and answers satisfaction questions, with no tasks being performed. The present research advanced the field by adopting the notion of user-centred translation (Suojanen, Koskinen and Tuominen 2015) and usability for final evaluation of machine translation. The UCT model defends the idea that the end-users' preference should be given priority as they should have the central role in the translation production. This research is inspired by their model and applies it to translation evaluation by using usability research and cognitive research in order to assess the usefulness of those translations as well as the end-

users' satisfaction with it. Moreover, this research moves from the previous mentioned MT evaluation studies by:

a) Implementing light post-editing rather than using only the raw machine translation output
b) Using lightly post-edited versions performed by professional translators who have experience in working with the industry partner and, therefore, are more aware of the content type
c) Testing the source content rather than just ignoring possible problems that could lead to problems in the translation as well
d) Implementing strong ecologically valid tasks to be performed by authentic users of the translation
e) Implementing a post-task questionnaire that, when answered right after the tasks are performed, are more likely to gather the real sentiment of satisfaction from users who performed the tasks
f) Implementing a web survey with genuine users of the software on a large scale
g) Making use of cognitive data in order to empirically test for usability
h) Implementing the methodology for challenging languages for MT systems
i) Creating a strong ecologically valid scenario since the steps for translation (MT, PE and HT), quality check and web survey followed the normal process any translation would in the company's everyday flow.

By implementing this strong approach for MT TQA, this research found that the impact that a translation has on the end user will vary according to the post-editing level and target language. This study found, in agreement with De Beaugrande and Dressler (1981) and Roturier (2006) that participants (from the German and Simplified Chinese groups, in the case of this study) found the translation *less* acceptable in its raw MT version since they were not able to tolerate the textual disturbances in the translation caused by the MT process. In contrast, German and Simplified Chinese participants found the translation much *more* acceptable in its lightly post-edited form as the disturbances in the text were solved by the post-editing process.

This study has shown that it is possible to evaluate the acceptability of machine-translated content via its different constituents: usability, quality and

satisfaction described in Chapter 3. As Chomsky (1969) and Puurtinen (1995) state, acceptability is a complex concept constructed of various degrees – which will depend on the purpose of the translation. In the case of this study, usability is one of these constituents. It is crucial to verify how usable the text is since the purpose of instructional texts is to offer instruction for specific tasks; satisfaction is another constituent since it informs us about the users' reactions and perceptions towards the translation when trying to perform the tasks; and quality was an additional constituent tested via accuracy and fluency in order to understand whether the quality ratings concur with the usability and satisfaction levels. This study successfully opens the field for more user-centred evaluation of machine translation.

8.3 Future Work

After successfully adapting the user-centered translation model by using usability methods for final machine translation evaluation, a further step would be a qualitative study on machine translation and post-editing errors in order to find whether the errors (if any) led to smaller degrees of acceptability. Another further step would be the implementation for a different set of languages, from more to less 'challenging' for machine translation. In addition, the need to measure acceptability for different content types is essential, as is the impact that this has on business factors such as willingness to buy or recommend a product or service, or even customer reputation. Several authors (Swales 1990; Shepherd and Waters 1998; Jiménez-Crespo 2013, among others) claim that content types, in particular web content, are constantly evolving and since they are characterised by their communicative purpose, content types are created, reshaped and extinguished by the user, based on the users' needs. Therefore, further investigation into the impact of different translation modes on specific content types, in different use scenarios, is warranted.

References

A

Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., González, J., Koehn, P., Leiva, L., Mesa-Lao, B., Ortiz, D., Saint-Amand, H., Sanchis, G. and Tsoukala, C. 2013. CASMACAT: An Open Source Workbench for Advanced Computer Aided Translation. *The Prague Bulletin of Mathematical Linguistics*, 100(1), pp.101–112.

Al-Qinai, J. 2000. Translation Quality Assessment: Strategies, Parametres and Procedures. *Meta* 45(3), pp.497–519.

Alt, F., Shirazi, A.S., Schmidt, A. and Mennenöh, J. 2012. Increasing the user's attention on the web: using implicit interaction based on gaze behavior to tailor content. *IN: Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, pp.544–553.

Alves, F., Pagano, A., Neumann, S., Steiner, E. and HansenSchirra, S. 2010. Translation Units and Grammatical Shifts: Towards an Integration of Product- and Process-based Translation Research. *IN:* Shreve, G. and Angelone, E. (eds.) *Translation and Cognition*. Amsterdam: John Benjamins, pp.109–142.

Alves, F., Gonçalves, J.L. and Szpak, K. 2012. Identifying instances of processing effort in translation through heat maps: An eye-tracking study using multiple input sources *IN:* Carl, M., Bhattacharya, P. and Choudhary, K.K. (eds.) *Proceedings of the First Workshop on Eye-tracking and Natural Language Processing at the 24th International Conference on Computational Linguistics*, pp.5–20.

Aziz, W., Sousa, S.C.M. and Specia, L. 2012. PET: a Tool for Post-editing and Assessing Machine Translation *IN:* Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. and Piperidis, S. (eds.) *Proceedings of the Eighth International Conference on Language Resources and Evaluation, 23-25 May 2012, Istanbul, Turkey*, pp.3982–3987.

Aziz, W., Koponen, M. and Specia, L. 2014. Sub-sentence Level Analysis of Machine Translation Post-editing Effort. *IN:* O'Brien, S., Balling, L.W., Carl, M., Simard, M. and Specia, L. (eds.) 2014. *Post-editing of Machine Translation: Processes and Applications*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp.170–199.

B

Baker, M. 1992. *In Other Words: A Coursebook on Translation*. London: Routledge.

Bevan, N., Kirakowski, J. and Maissel, J. 1991. What is Usability? *IN:* Bullinger, H.-J. (ed.) *Human Aspects in Computing, Design and Use of Interactive Systems and Work with Terminals: Proceedings of the Fourth International Conference on Human Computer Interaction*. Amsterdam: Elsevier, pp.651–655.

Birch, A., Osborne, M. and Koehn, P. 2008. Predicting success in machine translation. *IN: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp.745–754.

Bojar, O., Ercegovčević, M., Popel, M. and Zaidan, O.F. 2011. A grain of salt for the WMT manual evaluation. *IN: Proceedings of the 6th Workshop on Statistical Machine Translation, July 30-31, 2011, Edinburgh, Scotland*, pp.1–11.

Byrne, M.D., Anderson, J.R., Douglass, S. and Matessa, M. 1999. Eye tracking the visual search of click-down menus. *IN: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp.402–409.

Byrne, J. 2004. *Textual Cognetics and the Role of Iconic Linkage in Software User Guides*. PhD thesis, Dublin City University.

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.

C

Callison-Burch, C. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. *IN: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 6-7 August 2009, Singapore*, pp.286–295.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. and Schroeder, J. 2007. (Meta-)evaluation of machine translation. *IN: Proceedings of the Second Workshop on Statistical Machine Translation*, pp.136–158.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. and Schroeder, J. 2008. Further meta-evaluation of machine translation. *IN: Proceedings of the Third Workshop on Statistical Machine Translation*, pp.70–106.

Callison-Burch, C., Koehn, P., Monz, C. and Schroeder, J. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. *IN: Proceedings of the 4th EACL Workshop on Statistical Machine Translation, 30-31 March 2009, Athens, Greece*, pp.1–28.

Callison-Burch, C., Koehn, P., Monz, C. and Zaidan, O.F. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. *IN: Proceedings of the 6th Workshop on Statistical Machine Translation*, *July 30-31, 2011, Edinburgh, Scotland, UK*, pp.22–64.

Carl, M. 2012. Translog - II: a Program for Recording User Activity Data for Empirical Reading and Writing Research. *IN:* Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. and Piperidis, S. (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation, 23-25 May 2014, Istanbul, Turkey*, pp.4108–4112.

Carl, M. and Dragsted, B. 2012. Inside the Monitor Model: Processes of Default and Challenged Translation Production. *Translation: Computation, Corpora, Cognition* 2(1), pp.127–145.

Carl, M., Gutermuth, S. and Hansen-Schirra, S. 2015. Post-editing machine translation: A usability test for professional translation settings. *IN:* Ferreira, A. and Schwieter, J.W. (eds.) *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting*. Amsterdam: John Benjamins, pp.145–174.

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

Castilho, S. and O'Brien, S. 2016. *Content Profiling and Translation Scenarios. The Journal of Internationalization and Localization,* 3(1).

Catford, J. 1965. *A Linguistic Theory of Translation. Oxford: Oxford University Press.*

Chomsky, N. 1969. *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Correa, N. 2003. A Fine-grained Evaluation Framework for Machine Translation System Development. *IN: Proceedings of MT Summit IX. New Orleans, Louisiana, US.*

Coughlin, D. 2003. Correlating automated and human assessments of machine translation quality. *IN: Proceedings of the Machine Translation Summit IX, 23-27 September 2003, New Orleans, USA*, pp.63–70.

Crossley, S.A., Greenfield, J. and McNamara, D.S. 2008. Assessing Text Readability Using Cognitively Based Indices. *TESOL Quarterly*, 42(3), pp.475–493.

D

Daems, J., Macken, L. and Vandepitte, S. 2014. On the origin of errors: a fine-grained analysis of MT and PE errors and their relationship *IN:* Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. and Piperidis, S. (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation, 26-31 May 2014, Reykjavik, Iceland*, pp.62–66.

Daems, J., Vandepitte, S., Hartsuiker, R. and Macken, L. 2015. The Impact of Machine Translation Error Types on Post-Editing Effort Indicators. *IN: Proceedings of the 4th Workshop on Post-Editing Technology and Practice, November 3, 2015, Miami, USA*, pp.31–45.

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

De Beaugrande, R., and Dressier, W. 1981. *Introduction to Text Linguistics.* New York: Longman.

DePalma, D.A., Hegde, V., Pielmeier, H., Stewart, R.G. and Hedge, V. 2013. *The Language Services Market: 2013*. Lowell: Common Sense Advisory.

DePalma, D.A. and Sargent, B.B. 2013. *Transformative Translation: Machine Translation Will Change How Companies Provide Information to Global Customers*. Lowell: 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., Gaspari, F., Groves, D., van Genabith, J., Specia, L., Burchardt, A., Lommel, A. and Uszkoreit, H. 2013. *Mapping the Industry I: Findings on Translation Technologies and Quality Assessment* [Online]. Available from: http://www.qt21.eu/launchpad/sites/default/files/QTLP_Survey2i.pdf [Accessed 24 May 2016].

Doherty, S. and O'Brien, S. 2009. Can MT output be evaluated through eye tracking? *IN: Proceedings of the Twelfth Machine Translation Summit, August 26-30, 2009, Ottawa, Ontario, Canada*, pp.214–221.

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

Doherty, S. and O'Brien, S. 2012. A User-Based Usability Assessment of Raw Machine Translated Technical Instructions. *IN: Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas*, *28 October-1 November 2012, San Diego, California, USA*, pp.1–10.

Doherty, S. and 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), pp.40–51.

Dorr, B., Jordan, P. W. and Benoit, J. 1999. A survey of current paradigms in machine translation. *Advances in Computers*, 49(2), pp.1-68.

Drugan, J. 2013. *Quality in Professional Translation: Assessment and Improvement*. London: Bloomsbury Academic.

Duchowski, A.T. 2007. *Eye Tracking Methodology: Theory and Practice*. New York: Springer.

E

Ellis, S., Candrea, R., Misner, J., Craig, C.S., Lankford, C.P. and Hutchinson, T.E. 1998. Windows to the Soul? What Eye Movements Tell Us About Software Usability. *IN: Proceedings of the Usability Professionals' Association Conference*, pp.151–178.

F

Fields, P., Hague, D., Koby, G.S., Lommel, A. and Melby, A. 2014. What Is Quality? A Management Discipline and the Translation Industry Get Acquainted. *Revista Tradumàtica: tecnologies de la traducció* [Online], 12, pp.404–412. Available from: https://ddd.uab.cat/pub/tradumatica/tradumatica_a2014n12/tradumatica_a2014n12p404.pdf [Accessed 24 May 2016].

Flanagan, M. 1994. Error Classification for MT Evaluation. *IN: Proceedings of the Association of Machine Translation of the Americas, Washington, D.C,* pp.65-72.

Fuji, M. 1999. Evaluation Experiment for Reading Comprehension of Machine Translation Outputs. *IN: Proceedings of the Machine Translation Summit VII "MT in the Great Translation Era", 13-17 September 1999, Singapore*, pp.285–289.

Fuji, M., Hatanaka, N., Ito, E., Kamei, S., Kumai, H., Sukehiro, T., Yoshimi, T. and Isahara, H. 2001. Evaluation Method for Determining Groups of Users Who Find MT "Useful". *IN: Proceedings of the Machine Translation Summit VIII "Machine Translation in the Information Age", 18-22 September 2001, Santiago de Compostela, Spain*, pp.103–108.

G

Gaspari, F. 2004. Online MT Services and Real Users' Needs: An Empirical Usability Evaluation *IN:* Frederking, R.E. and Taylor, K.B. (eds.) *Proceedings of AMTA 2004: 6th Conference of the Association for Machine Translation in the Americas "Machine Translation: From Real Users to Research"*. Berlin: Springer, pp.74–85.

Gaspari, F., Toral, A., Lommel, A., Doherty, S., van Genabith, J. and Way, A. 2014. Relating Translation Quality Barriers to Source-Text Properties. *IN: Proceedings of the Workshop on Automatic and Manual Metrics for Operational Translation Evaluation at LREC 2014, 26 May 2014, Reykjavik, Iceland*, pp.61–70.

Giménez, J. and Màrquez, L. 2008. A smorgasbord of features for automatic MT evaluation. *IN: Proceedings of the Third Workshop on Statistical Machine Translation*, pp.195–198.

Göpferich, S., Jakobsen, A.L. and Mees, I.M. (eds.) 2008. *Looking at Eyes: Eye-Tracking Studies of Reading and Translation Processing*. Frederiksberg: Samfundslitteratur.

Giménez, J., Màrquez, L., Comelles, E., Catellón, I. and Arranz, V. 2010. Document-level Automatic MT Evaluation based on Discourse Representations. *IN: The Joint Fifth*

Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, Sweden, pp.333–338.

Göpferich, S., Alves, F. and Mees, I.M. (eds.) 2010. *New Approaches in Translation Process Research*. Frederiksberg: Samfundslitteratur.

Görög, A. 2014. Quantifying and benchmarking quality: the TAUS Dynamic Quality Framework. *Revista Tradumàtica: tecnologies de la traducció* [Online], 12, pp.404–412. Available from: http://ddd.uab.cat/pub/tradumatica/tradumatica_a2014n12/tradumatica_a2014n12p443.pdf [Accessed 02 June 2016].

Goto, S., Lin, D. and Ishida, T. 2014. Crowdsourcing for Evaluating Machine Translation Quality *IN:* Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. and Piperidis, S. (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation, 26-31 May 2014, Reykjavik, Iceland*, pp.3456–3463.

Graesser, A.C., McNamara, D.S., Louwerse, M.M. and Cai, Z. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), pp.193–202.

Graham, Y., Baldwin, T., Moffat, A. and Zobel, J. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. *IN: Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, August 8-9, 2013, Sofia, Bulgaria*, pp.33–41.

Graham, Y., Baldwin, T., Moffat, A. and Zobel, J. 2015. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView, pp.1–28.

Guerberof, A.A. 2014. Correlations Between Productivity and Quality when Postediting in a Professional Context. *Machine Translation*, 28(3-4), pp.165-186.

Guzmán, F., Joty, S., Màrquez, L. and Nakov, P. 2014. Using Discourse Structure Improves Machine Translation Evaluation. *IN: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, June 23-25 2014, Baltimore, Maryland, USA,* pp.687–698.

H

Holmes, J.S. 1988. *Translated! Papers on Literary Translation and Translation Studies*. Amsterdam: Rodopi.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H. and van de Weijer, J. (eds.) 2011. *Eye Tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.

House, J. 1997. *Translation Quality Assessment. A Model Revisited*. Tübingen: Gunter Narr.

House, J. 2015. *Translation Quality Assessment: Past and Present*. London: Routledge.

Hovy, E., King, M. and Popescu-Belis, A. 2002. Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, 17(1), pp.43–75.

Howitt, D. and Cramer, D., 2005. An Introduction to Statistics in Psychology. Pearson Education.

Howitt, D. and Cramer, D. 2011. *Introduction to Statistics in Psychology*. Harlow: Prentice Hall.

Hvelplund, K.T. 2011. *Allocation of Cognitive Resources in Translation: An Eye-tracking and Key-logging Study*. PhD thesis, Copenhagen Business School.

I

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/TS 16982:2002 *Ergonomics of human-system interaction – Usability methods supporting human-centred design.* Geneva: International Organization for Standardization.

ISO 2010. ISO/TS 11669:2012 *Technical Specification: Translation projects — General guidance*. Geneva: International Organization for Standardization.

J

Jacob, R.J.K. 1995. Eye tracking in advanced interface design *IN:* Barfield, W. and Furness, T.A. (eds.) *Virtual Environments and Advanced Interface Design*. Oxford: Oxford University Press, pp.258–288.

Jacob, R.J.K. and Karn, K.S. 2003. Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises *IN:* Hyönä, J., Radach, R. and Deubel, H. (eds.) *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*. Amsterdam: Elsevier, pp.573–605.

Jakobsen, A.L. and Jensen K.T.H. 2008. Eye movement behaviour across four different types of reading task. *IN:* Göpferich, S., Jakobsen, A.L. and Mees, I.M. (eds.) 2008. *Looking at Eyes: Eye-Tracking Studies of Reading and Translation Processing*. Frederiksberg: Samfundslitteratur, pp.103–124.

Jensen, K.T.H. 2009. Indicators of text complexity. *IN:* Göpferich, S., Jakobsen, A.L. and Mees, I.M. (eds.) 2009. *Behind the Mind: Methods, Models and Results in Translation ProcessResearch*. (Copenhagen Studies in Language 36). Copenhagen: Samfundslitteratur, pp.61-80.

Jimenez-Crespo, M.A. 2013. *Translation and Web Localization*. Routledge.

Johnson, R.R., Salvo, M.J. and Zoetewey, M.W. 2007. User-Centered Technology in Participatory Culture: Two Decades "Beyond a Narrow Conception of Usability Testing". *IEEE Transactions on Professional Communication*, 50(4), pp.320–332.

Jones, D., Gibson, E., Shen, W., Granoien, N., Herzog, M., Reynolds, D. and Weinstein, C. 2005. Measuring Human Readability Of Machine Generated Text: Three Case Studies In Speech Recognition And Machine Translation. *IN: Proceedings of ICASSP '05 IEEE International Conference on Acoustics, Speech, and Signal Processing 2005 – Volume 5, 18-23 March 2005, Philadelphia, USA*, pp.1009–1012.

K

Karn, K.S., Ellis, S. and Juliano, C. 1999. The Hunt for Usability: Tracking Eye Movements. *IN: Extended Abstracts on Human Factors in Computing Systems*. New York: ACM, p.173.

Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T. and Yoon, H.-J. 2015. Eye-tracking analysis of user behavior and performance in web search on large and small screens. *Journal of the Association for Information Science and Technology*, 66(3), pp.526–544.

Klerke, S., Castilho, S., Barrett, M. and Søgaard, A. 2015. Reading metrics for estimating task efficiency with MT output. *IN: Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning, 18 September 2015, Lisbon, Portugal,* pp.6–13.

Kliegl, R., Grabner, E., Rolfs, M. and Engbert, R. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16(1/2), pp.262–284.

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

Koehn, P. 2010. Enabling Monolingual Translators: Post-Editing vs. Options. *IN: Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, *Los Angeles, California,* pp.537–545.

Koponen, M. 2012. Comparing Human Perceptions of Post-Editing Effort with Post-Editing Operations. *IN: Proceedings of the Seventh Workshop on Statistical Machine Translation, June 7-8, 2012, Montréal, Canada,* pp.181–190.

Koponen, M. 2016. Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. *The Journal of Specialised Translation* [Online], 25, pp.131–148. Available from: http://www.jostrans.org/issue25/art_koponen.pdf [Accessed 24 May 2016].

Krings, H. P. 2001. *Repairing texts: empirical investigations of machine translation postediting processes.* Kent, OH, USA: The Kent State University Press, edited/translated by G. S. Koby.

L

Lacruz, I and Shreve, G. M. 2014. Pauses and cognitive effort in post-editing. *IN:* O'Brien, S., Balling, L.W., Carl, M., Simard, M. and Specia, L. (eds.) 2014. *Post-editing of Machine Translation: Processes and Applications*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp.246-272.

Larose, R. 1987. *Théories contemporaines de la traduction.* Quebec: Presses de l'Université du Québec, 2nd edition.

Lassen, I. 2003. *Accessibility and Acceptability in Technical Manuals: A survey of style and grammatical metaphor*. Amsterdam: John Benjamins.

Lauscher, S. 2000. Translation Quality Assessment: Where Can Theory and Practice Meet? *The Translator*, 6(2), pp.149–168.

Lavie, A. and Agarwal, A. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. *IN: Proceedings of the Workshop on Statistical Machine Translation, June, Prague, Czech Republic,* pp.228–231 .

LDC, 2002. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5.

Liu, C., Dahlmeier, D., and Ng, H.T. 2011. Better Evaluation Metrics Lead to Better Machine Translation. *IN: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, July 27-31, 2011, Edinburgh, Scotland, UK*, pp.375–384.

Litjós, A.F, Carbonell, J.G. and Lavie, A. 2005. A Framework for Interactive and Automatic Refinement of Transfer-based Machine Translation. *IN: Proceedings of the Tenth Annual Conference of the European Association for Machine Translation, 30-31 May 2005, Budapest, Hungary*, pp.87-96.

Lommel, A., Uszkoreit, H. and Burchardt, A. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista Tradumàtica: tecnologies de la traducció* [Online], 12, pp.455–463. Available from: https://ddd.uab.cat/pub/tradumatica/tradumatica_a2014n12/tradumatica_a2014n12p455.pdf [Accessed 24 May 2016].

M

McCarthy, P.M., Lewis, G.A., Dufty, D.F. and McNamara, D.S. 2006. Analyzing Writing Styles with Coh-Metrix. *IN: Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS)*, pp.764–769.

McNamara, D.S., Graesser, A.C., McCarthy, P.M. and Cai, Z. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

Mitchell, L. 2015. *Community post-editing of machine-translated user-generated content*. PhD thesis, Dublin City University.

Mitchell, L., O'Brien, S. and Roturier, J. 2014. Quality evaluation in community post-editing. *Machine Translation*, 28(3), pp.237–262.

Molnár, O. 2012. Source Text Quality in the Translation Process *IN:* Zehnalová, J., Molnár, O. and Kubánek, M. (eds.) *Tradition and Trends in Trans-Language Communication*. Olomouc: Palacký University, pp.59–86.

Moravcsik, J. and Kintsch, W. 1995. Writing quality, reading skills, and domain knowledge as factors in text comprehension. *IN:* Henderson, J., Singer, M. and Ferreira, F. (eds.) 1995. *Reading and Language Processing.* New York, London: Psychology Press, pp.232-246.

Moorkens, J. 2012. A mixed-methods study of consistency in translation memories. *Localisation Focus*, 11(1), pp.14–26.

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

Moran, J., Lewis, D. and Saam, C. 2014. Analysis of post-editing data: a productivity field test using and instrumented CAT tool. *IN:* O'Brien, S., Balling, L.W., Carl, M., Simard, M. and Specia, L. (eds.) 2014. *Post-editing of Machine Translation: Processes and Applications*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp.128-169.

Muegge, U. 2013. Do translation standards encourage effective terminology management? *Revista Tradumàtica: tecnologies de la traducció* [Online], 13, pp.552–560. Available from: https://ddd.uab.cat/pub/tradumatica/tradumatica_a2015n13/tradumatica_a2015n13p552.pdf [Accessed 24 May 2016].

Munday, J. 2008. *Introducing Translation Studies: Theories and Applications*. London: Routledge.

Muzii, L. 2014. The red-pen sindrome. *Revista Tradumàtica: tecnologies de la traducció* [Online], 12, pp.421–429. Available from:

https://ddd.uab.cat/pub/tradumatica/tradumatica_a2014n12/tradumatica_a2014n12p421.pdf [Accessed 24 May 2016].

N

Nida, E. 1964. *Toward a Science of Translation*. Leiden: Brill.

Nielsen, J. 1993. *Usability Engineering*. Amsterdam: Morgan Kaufmann.

Nielsen, J. 2006. F-Shaped Pattern For Reading Web Content [Online]. April 17, 2006. Available from: https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/ [Accessed 23 January 2016].

Nießen, S., Och, F.J., Leusch, G. and Ney, H. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. *IN: Proceedings of the Second International Conference on Language Resources and Evaluation, 31 May-2 June 2000, Athens, Greece,* pp.39–45.

Nord, C. 2014 Quo vadis, functional translatology? *IN:* Brems, E., Meylaerts, R. and van Doorslaer, L. (eds.) *The Known Unknowns of Translation Studies*. Amsterdam: John Benjamins, pp.29–45.

O

O'Brien, S. 2006. Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology*, 14(3), pp.185–205.

O'Brien, S. 2009. Eye tracking in translation-process research: methodological challenges and solutions *IN:* Mees, I.M., Alves, F. and Göpferich, S. (eds.) *Methodology, Technology and Innovation in Translation Process Research: A Tribute to Arnt Lykke Jakobsen*. Frederiksberg: Samfundslitteratur, pp.251–266.

O'Brien, S., Roturier, J. and de Almeida, G. 2009. *Post-Editing MT Output: Views from the researcher, trainer, publisher and practitioner.* Tutorial given at the *Machine Translation Summit XII, August 26, 2009, Ottawa, Ontario, Canada*.

O'Brien, S., 2010. Controlled language and readability *IN:* Shreve, G.M. and Angelone, E. (eds.) *Translation and cognition*. Amsterdam: John Benjamins, pp.143–165.

O'Brien, S. 2011. Towards predicting post-editing productivity. *Machine Translation*, 25(3), pp.197–215.

O'Brien, S., Choudhury, R., Van der Meer, J. and Aranberri Monasterio, N. 2011. *Dynamic Quality Evaluation Framework – November 2011* [Online]. De Rijp: TAUS. Available from: https://goo.gl/eyk3Xf [Accessed 26 May 2016].

O'Brien, S. 2012. Towards a Dynamic Quality Evaluation Model for Translation. *The Journal of Specialised Translation* [Online], 17, pp.55–77. Available from: http://www.jostrans.org/issue17/art_obrien.pdf [Accessed 02 May 2016].

O'Brien, S., Simard, M. and 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, October 28.

O'Brien, S., Simard, M. and Specia, L. (eds.) 2013. Workshop on Post-editing Technology and Practice (WPTP 2013). *Machine Translation Summit XIV.* Nice, September 2-6.

O'Brien, S., Balling, L.W., Carl, M., Simard, M. and Specia, L. (eds.) 2014. *Post-editing of Machine Translation: Processes and Applications*. Newcastle upon Tyne: Cambridge Scholars Publishing.

O'Hagan, M. 2011. Community Translation: Translation as a Social Activity and Its Possible Consequences in the Advent of Web 2.0 and Beyond *IN:* O'Hagan, M. (ed.) *Linguistica Antverpiensia: Special Issue on Translation as a Social Activity*, 10, pp.11–23.

P

Padó, S., Cer, D., Galley, M., Jurafsky, D. and Manning, C. D. 2009. Measuring Machine Translation Quality as Semantic Equivalence: A Metric Based on Entailment Features. *Machine Translation,* 23(2-3), pp.181–193.

Papineni, K. Roukos, S. Ward, T. and Zhu, W. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. *IN: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania,* pp.311–318,

Plitt, M. and F. Masselot. 2010. A productivity test of statistical machine translation postediting in a typical localisation context. *IN: The Prague Bulletin of Mathematical Linguistics.* Prague, Czech Republic: Universita Karlova, pp. 7-16.

Puurtinen, T. 1995. *Linguistic acceptability in translated children's literature*. PhD thesis, University of Joensuu.

Pym, A. 2010. *Exploring Translation Theories*. Abingdon: Routledge.

R

Radach, R., Kennedy, A. and Rayner, K. (eds.) 2004. *Eye Movements and Information Processing During Reading. A Special Issue of The European Journal of Cognitive Psychology*. Hove: Psychology Press.

Rayner, K. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, 124(3), pp.372–422.

Rayner, K. and Juhasz, B. J. 2004. Eye movements in reading: Old questions and new directions. *European Journal of Cognitive Psychology*, 16, pp.340–352.

Reiss, K. 1971. *Möglichkeiten und Grenzen der Übersetzungskritik*. Munich: Hueber.

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*. PhD thesis, Dublin city University.

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

S

Schäffner, C. 1997. From 'Good' to 'Functionally Appropriate': Assessing Translation Quality. *Current Issues in Language and Society,* 4(1), pp.1-5.

Secară, A. 2005. Translation Evaluation: A State of the Art Survey. *IN: Proceedings of the eCoLoRe/MeLLANGE Workshop, 21-23 March 2005, Leeds, UK,* pp.39–44.

Shrestha, S. and Lenz, K. 2007. Eye Gaze Patterns while Searching vs. Browsing a Web site. *Usability News* [Online], 14 January, 9(1). Available from: http://usabilitynews.org/eye-gaze-patterns-while-searching-vs-browsing-a-website/ [Accessed 26 May 2016].

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. 2006. A study of translation edit rate with targeted human annotation. *IN: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: "Visions for the Future of Machine Translation", August 8-12, 2006, Cambridge, Massachusetts, USA*, pp.223–231.

Sousa, S.C., Aziz, W., Specia, L., 2011. Assessing the Post-Editing Effort for Automatic and Semi-Automatic Translations of DVD Subtitles., in: RANLP. Presented at the Recent Advances in Natural Language Processing, pp.97–103.

Specia, Lucia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. *IN: Proceedings of the Fifteenth Annual Conference of the European Association for Machine Translation, 30-31 May,* Leuven, Belgium, pp.73–80

Starr, M.S. and Rayner, K. 2001. Eye movements during reading: some current controversies. *TRENDS in Cognitive Sciences*, 5(4), pp.156–163.

Stymne, S. and Ahrenberg, L. 2012. On the practice of error analysis for machine translation evaluation *IN:* Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. and Piperidis, S. (eds.) *Proceedings of the Eighth International Conference on Language Resources and Evaluation, 23-25 May 2012, Istanbul, Turkey*, pp.1785–1790.

Stymne, S., Danielsson, H., Bremin, S., Hu, H., Karlsson, J., Lillkull, A.P. and Wester, M. 2012. Eye Tracking as a Tool for Machine Translation Error Analysis *IN:* Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. and Piperidis, S. (eds.) *Proceedings of the Eighth International Conference on Language Resources and Evaluation, 23-25 May 2012, Istanbul, Turkey*, pp.1121–1126.

Suojanen, T., Koskinen, K. and Tuominen, T. 2015. *User-Centered Translation*. Abingdon: Routledge.

Suokas, J., Pukarinen, K., von Wolff, S. and Koskinen, K. 2015. Testing Testing: Putting Translation Usability to the Test. *Journal of Translation and Technical Communication Research*, 8(2), pp.499–519.

Swales, J. 1990. *Genre analysis: English in academic and research settings.* Cambridge, UK:Cambridge University Press.

T

Tatsumi, M. 2009. Correlation between automatic evaluation scores, post-editing speed and some other factors. *IN: Proceedings of MT Summit XII, Ottawa, 26–30 August 2009,* pp.332–339.

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

Toury, G. 1995. *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.

U

Uszkoreit, H. and Lommel, A. 2013. Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment. Paper presented at *Localization World, 12-14 June 2013, London, United Kingdom*.

V

Vandepitte, S., Maylath, B., Mousten, B., Minacori, P. and Scarpa, F. 2010. Interactivities between professional translators and professional communicators: what translators would like communicators to know. *IN: Proceedings of the 2010 IEEE International Professional Comunication Conference*, *7-9 July 2010, Enschede, Netherlands*, pp.58–59.

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

Vieira, L.N. 2014. Indices of cognitive effort in machine translation post-editing. *Machine Translation*, 28(3), pp.187–216.

Vilar, D., Xu, J., D'Haro, L.F. and Ney, H. 2006. Error Analysis of Statistical Machine Translation Output. *IN: Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation, 22-28 May 2006, Genoa, Italy,* pp.697–702.

W

Watters, C., and Shepherd, M. A. 1997. The digital broadsheet: An evolving genre. *IN: Proceedings of the Thirtieth Hawaii International Conference on SystemSciences* 6, pp. 22-29.

White, J.S., O'Connell, T.A. and Carlson, L.M. 1993. Evaluation of machine translation. *IN: Proceedings of the Workshop on Human Language Technology, March 21-24, 1993, Plainsboro, New Jersey, USA*. San Francisco: Morgan Kaufmann, pp.206–210.

White, J.S. and O'Connell, T.A. 1994. Evaluation in the ARPA machine translation program: 1993 methodology. *IN: Proceedings of the Workshop on Human Language Technology, March 8-11, 1994, Plainsboro, New Jersey, USA.* San Mateo: Morgan Kaufmann, pp.134–140.

White, J.S., O'Connell, T. and O'Mara, F. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. *IN: Technology partnerships for crossing the language barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas,5-8 October 1994, Columbia, Maryland, USA*, pp.193–205.

White, J.S. and O'Connell, T.A. 1996. Adaptation of the DARPA machine translation evaluation paradigm to end-to-end systems. *IN: Expanding MT Horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas, 2-5 October 1996, Montreal, Quebec, Canada*, pp.106–114.

Williams, M. 2004. *Translation Quality Assessment: An Argument-centered Approach*. Ottawa: University of Ottawa Press.

Williams, R. and Morris, R. 2004. Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology*, 16(1-2), pp.312–339.

Williams, J. 2013. *Theories of Translation*. Basingstoke: Palgrave Macmillan.

Wong, B.T.M. and Kit, C. 2012. Extending Machine Translation Evaluation Metrics with Lexical Cohesion to Document Level. *IN: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 12-14 July 2012, Jeju Island, Korea*, pp.1060–1068.

Z

Zaidan, O.F. and Callison-Burch, C. 2011. Crowdsourcing Translation: Professional Quality from Non-Professionals. *IN: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, June 19-24, 2011, Portland, Oregon*, pp.1220–1229.

Appendices

Contents

APPENDIX A: Informed Consent Form and Plain Language Statement

DUBLIN CITY UNIVERSITY

Plain Language Statement

I. Introduction to the Research Study

Human interaction with MT output: Usability, Acceptability, Post-editing Research

Dr. Sharon O'Brien, School of Applied Language and Intercultural Studies, Dublin City University
t. (01) 7005381
e. sharon.obrien@dcu.ie

Sheila Castilho M de Sousa, School of Applied Language and Intercultural Studies, CNGL, Dublin City University
e. sheila.castilhomdesousa3@mail.dcu.ie

II. Details of what involvement in the Research Study will require

For the interview:

The requirement for participation in this project is to read the Plain Language Statement, fill in and return the Informed Consent Form, and then answer some questions regarding the types of contents the company generally translates using HT, MT and PE. The interview is expected to last around 30 minutes.

For the eye tracking:

The requirement for participation in this project is to read the Plain Language Statement, fill in and return the Informed Consent Form, and then answer some pre-task questions. You will then be asked to read some content and perform a task using a piece of software on the computer. Once the task is completed, you may be asked some post-task questions, which may also involve measuring recall and comprehension of the content you read.

For the online survey:

The requirement for participation in this project is to read the Plain Language Statement, fill in and return the Informed Consent Form, and then answer some questions online on the quality of translated content that will be presented to you in the survey.

III. Potential risks to participants from involvement in the Research Study (if greater than that encountered in everyday life)

There are no risks to participants.

IV. Benefits (direct or indirect) to participants from involvement in the Research Study

The research focuses on the human interaction with machine translated texts, as well as human translation and post-edited texts.
Our focus is on identifying how usable and acceptable those contents are for the end-user. Usability and acceptability metrics will aid automatic routing of Enterprise content. This will have benefits both for end users of translated content and for Enterprises who engage in translation.

V. Advice as to arrangements to be made to protect confidentiality of data, including that confidentiality of information provided is subject to legal limitations

No identifying personal data will be requested. Only researchers involved in the project will have access to the responses. All data will be stored securely on password protected PCs at DCU.

VII. Statement that involvement in the Research Study is voluntary

Involvement in this study is completely voluntary and participants may withdraw from the study at any point.

VIII. Any other relevant information

For any participant who may have an existing relationship with Dublin City University, involvement/non-involvement in this project will not affect your ongoing relationship with Dublin City University in any way.

If participants have concerns about this study and wish to contact an independent person,

please contact:

The Secretary, Dublin City University Research Ethics Committee, c/o Research and Innovation Support, Dublin City University, Dublin 9. Tel 01-7008000

Informed Consent Form

I. Research Study Title

Human interaction with MT output: Usability, Acceptability, Post-editing Research

Dr. Sharon O'Brien, School of Applied Language and Intercultural Studies, Dublin City University

Sheila Castilho M de Sousa, School of Applied Language and Intercultural Studies, CNGL, Dublin City University

II. Clarification of the purpose of the research

Usability and acceptability of translated and un-translated content can impact on brand image, customer loyalty, and sales. Different stakeholders will have different acceptability thresholds; some will want high quality, others may make do with faster turnaround, lower quality, or might even prefer non-translated content compared with raw Machine Translation (PEZero).

The first phase of the research will involve interviewing Translation Project Managers from our partners companies which will help us to understand what kind of contents are Human Translated or Machine Translated and how the process is decided. We will report profiling translation scenarios within the enterprise focusing on scenarios where there is no translation, raw Machine Translation, and Computer-Aided Translation, with human post-editing and decisions that push content in one or another direction.

Subsequently, we intend to recruit volunteers and make use of the eye-tracker and a post-task questionnaire to understand how usable and acceptable the translation and non-translation scenarios are for the end-user.

III. Confirmation of particular requirements as highlighted in the Plain Language Statement

The requirement for participation in this project is to read the Plain Language Statement, fill in and return the Informed Consent Form and then, if being interviewed, answer some questions for the interview.

If you are partaking in the eye tracking aspect of the study, the requirement for participation is to read the Plain Language Statement, fill in and return the Informed Consent Form, and then answer some pre-task questions. You will then be asked to read some content and perform a task using a piece of software on the computer. Once the task is completed, you may be asked some post-task questions, which may also involve measuring recall and comprehension of the content you read.

If you are participating in the online survey, the requirement for participation is to read the Plain Language Statement, fill in and return the Informed Consent Form, and then answer some questions online on the quality of translated content that will be presented to you in the survey.

Participant – please complete the following (Yes or No for each question)

I have read the Plain Language Statement (or had it read to me)	Yes/No
I understand the information provided	Yes/No
I have had an opportunity to ask questions and discuss this study	Yes/No
I have received satisfactory answers to all my questions	Yes/No
I am aware that my interview will be audiotaped	Yes/No

IV. Confirmation that involvement in the Research Study is voluntary

I may withdraw from the Research Study at any point.

V. Advice as to arrangements to be made to protect confidentiality of data, including that confidentiality of information provided is subject to legal limitations

No personal identification details will be requested in the interview and eye-tracker tasks. Only the researchers involved in the study will have access to the questionnaire responses.

VI. Any other relevant information

For any participant who may have an existing relationship with Dublin City University, involvement/non-involvement in this project will not affect your ongoing relationship with Dublin City University in any way.

VII. Signature:

I have read and understood the information in this form. My questions and concerns have been answered by the researchers, and I have a copy of this consent form. Therefore, I consent to take part in this research project

Participants Signature: _____

Name in Block Capitals: _____

Witness: _____

Date: _____

APPENDIX B: Tasks

English:

Open the first tab in the Excel spread sheet called "Calendar" and perform tasks from 1 to 5

1) Quickly change colors, fonts, and effects in your worksheet

a) To switch to another theme, click Page Layout > Themes, and pick "Facet".

b) Click Page Layout > Fonts, and pick "Office".

2) Change the font format for hyperlinks

a) Click the cell with the hyperlink. On the Home tab, right-click the Hyperlink style and pick Modify.

b) In the Style box, click Format.

c) Click Font, choose "Arial Black" and click OK.

d) Click OK to close the Style box.

3) Format text in headers or footers

a) On the status bar, click the Page Layout View button. 

b) Select the header text.

c) On the Home tab in the Font group, pick "Arial Black".

d) When you're done, click the Normal view button on the status bar. 

4) Add a comment

a) Select the cell A1 to add a comment to and do one of the following:

   1. On the Review tab, in the Comments group, click New Comment.
   2. Right-click the cell and then click Insert Comment.
   3. A new comment is created, and the pointer moves to the comment. An indicator appears in the corner of the cell.

b) In the body of the comment, type the comment text "ok".

c) Click outside the comment box.

d) The comment box disappears, but the comment indicator remains.

5) Apply conditional formatting with color in Excel

a) Pick the column A where you want to format duplicate values with a color
b) On the Home tab, click Conditional Formatting > Highlight Cells Rules > Text that contains…
c) Type PUBLIC HOLIDAY
d) Choose Green Fill with Dark Green Text
e) Click OK to format the cells.
f) The duplicate values are highlighted with a light green fill and dark green text.

Now, open the tab called "Calculate" and perform tasks 6-8

6) Insert an exploding pie chart

1. In your spreadsheet, select the data to use for your pie chart (A3 to B8).
2. Click Insert > Insert Pie.
3. Under 2-D Pie, choose the leftmost option, Pie.
4. To explode the pie chart, do the following:
   a) Click the chart, then select the whole pie.
   b) Under Chart Tools, click the Format tab and then click Format Selection.
   c) In the Format Data Series pane, change the percentage value in the Pie Explosion box, under Series Option, to explode the pie. Set it to 30%.
5. To add a title to your chart, select the chart, pick the Chart Elements button, and then check the Chart Title box.
6. If there's already a title, such as "Chart Title," replace it by typing "Dates" in the title box
7. Under Chart Tools, click Design > Add Chart Element.
8. Point to Data Labels, and click More Data Label Options.
9. In the Format Data Labels pane on the right, click Label Options.
10. Uncheck the Value box and check the Percentage box.
11. Click  , click Fill, and pick the Gradient fill button.

7) Insert a bar of pie chart

a) In your spreadsheet, select your pie chart.
b) Click Insert > Insert Pie.

c) Under 2-D Pie, choose the rightmost option, Bar of Pie.
d) To change the colors that the chart uses, click the Chart Styles button, and click Color.
e) Pick a color theme under Colorful or Monochromatic, such as Color 4.

8) Hide comments and their indicators

a) Click the File tab, then click Options.
b) In the Advanced category, under Display, do the following:
   i. To hide both comments and indicators throughout the workbook, under For cells with comments, show, click No comments or indicators.

German:

MT

Open the first tab in the Excel spread sheet called "Calendar" and perform tasks from 1 to 5

1) Schnell zu ändern, Farben, Schriftarten und Effekte auf einem Arbeitsblatt

c) Wenn Sie in ein anderes Design wechseln möchten, klicken Sie auf Seitenlayout > Designs, und wählen Sie "Facette".
d) Klicken Sie auf Seitenlayout > Schriftarten, und wählen Sie "Office".

2) Ändern des Formats der Schriftart für hyperlinks

e) Klicken Sie auf die Zelle mit dem Hyperlink. Auf der **Start** Registerkarte der rechten Maustaste auf die **Hyperlink** Stil aus, und wählen Sie **Ändern**.
f) In der **Stil** auf **Format**.
g) Klicken Sie auf **Schriftart**, wählen Sie "Arial Black", und klicken Sie auf **OK**.
h) Klicken Sie auf **OK** zum Schließen der **Stil** Feld.

3) Formatieren von Text in Kopf- oder Fußzeilen

e) Klicken Sie auf der Statusleiste auf die **Seitenlayoutansicht** Schaltfläche.

f) Wählen Sie Text in die Kopfzeile.

g) Auf der **Start** Registerkarte die **Schriftart** Gruppe, wählen Sie "Arial Black" aus.

h) Wenn Sie fertig sind, klicken Sie auf die **Normal** Schaltfläche auf der Statusleiste angezeigt. 

4) Hinzufügen von Kommentaren

a. Klicken Sie auf die Zelle A1 und führen Sie eine der folgenden Aktionen aus:

1. Klicken Sie auf der Registerkarte Überprüfen in der Gruppe Kommentare auf Neuer Kommentar.
2. Klicken Sie mit der rechten Maustaste auf die Zelle, und klicken Sie dann auf **Kommentar einfügen**.
3. Es wird ein neuer Kommentar erstellt und der Zeiger zum Kommentar bewegt. In der Ecke der Zelle wird ein Indikator angezeigt.

b) Geben Sie im Nachrichtenteil des Kommentars den Kommentartext ein “**ok**”.

c) Klicken Sie außerhalb des Kommentarfelds.

d) Das Kommentarfeld wird nicht mehr angezeigt, aber der Kommentarindikator bleibt erhalten.

5) Anwenden von bedingter Formatierung in Excel farbig

g) Wählen Sie die Spalte A, wo Sie die doppelten Werte mit einer Farbe zu formatieren möchten

h) Auf der **Start** auf **bedingte Formatierung** > **Regeln zum Hervorheben von Zellen > Text,... enthält**

i) Typ **"PUBLIC HOLIDAY"**

j) Wählen Sie **grüner Füllung mit dunkelgrünem Text**

k) Klicken Sie auf **OK** um Zellen zu formatieren.

l) Die doppelten Werte werden mit einem helle grüner Füllung und mit dunkelgrünem Text hervorgehoben.

Now, open the tab called “Calculate” and perform tasks 6-8

6) Einfügen eines Kreisdiagramms explodieren

12. Wählen Sie in der Kalkulationstabelle, die Daten für das Kreisdiagramm verwenden (A3 B8).

13. Klicken Sie auf **Einfügen** > **Kreis einfügen**.
14. Klicken Sie unter **2D-Kreisdiagramm**, wählen Sie die am weitesten links stehende Option **Kreis**.
15. Um das Kreisdiagramm zu entfalten, führen Sie folgende Schritte aus:

d) Klicken Sie auf das Diagramm, und wählen Sie dann des gesamten Kreises.

e) Klicken Sie unter **Diagrammtools**, klicken Sie auf die **Format** Registerkarte, und klicken Sie dann auf **Formatauswahl**.

f) In der **Datenreihen formatieren** Bereich, ändern Sie den Wert der Prozentsatz

in der **Kreisexplosion** im Feld **Option Reihe**, um die Explosion des Kreises. Legen Sie es auf 30 %.

16. Wenn Sie einen Titel zu Ihrem Diagramm hinzufügen möchten, wählen Sie das Diagramm, wählen Sie aus der **Diagrammelemente** Schaltfläche, und aktivieren Sie dann die **Diagrammtitel** Feld.
17. Wenn es bereits ein Titel ein, z. B. "Diagrammtitel", ersetzen Sie ihn durch "Datum" in das Titelfeld
18. Klicken Sie unter **Diagrammtools**, klicken Sie auf **Design** > **Diagrammelement hinzufügen**.
19. Zeigen Sie auf **Datenbeschriftungen**, und klicken Sie auf **Weitere Datenbeschriftungsoptionen**.
20. In der **Datenbeschriftungen formatieren** Bereich rechts auf **Beschriftungsoptionen**.
21. Deaktivieren Sie die **Wert** ein, und überprüfen Sie die **Prozentsatz** Feld.
22. Klicken Sie auf  , klicken Sie auf **Füllung**, und wählen Sie die **Farbverlauf** Schaltfläche.

7) Einfügen eines Balkens-aus-Kreis-Diagramm

f) Wählen Sie in der Kalkulationstabelle das Kreisdiagramm ein.

g) Klicken Sie auf **Einfügen** > **Kreis einfügen**.

h) Klicken Sie unter **2D-Kreisdiagramm**, wählen Sie die Option äußerst **Balken aus Kreis**.

i) Zum Ändern der Farben, die im Diagramm verwendet werden, klicken Sie auf die **Diagrammformatvorlagen** Schaltfläche, und klicken Sie auf **Farbe**.

j) Auswählen eines Designs "Farbe" unter **farbig** oder **"Monochromatisch"**, z. B. Farbe 4.

8) Ausblenden von Kommentaren und ihren Indikatoren

a) Klicken Sie auf die Registerkarte Datei und dann auf Optionen.

b) Führen Sie in der Kategorie Erweitert unter Anzeige eine der folgenden Aktionen aus:

   i. Damit sowohl Kommentare als auch Indikatoren in der gesamten Arbeitsmappe ausgeblendet werden, klicken Sie unter Für Zellen mit Kommentaren Folgendes anzeigen auf Keine Kommentare und Indikatoren.

PE

Open the first tab in the Excel spread sheet called "Calendar" and perform tasks from 1 to 5

1) Farben, Schriftarten und Effekte in einem Arbeitsblatt schnell ändern

a) Wenn Sie das Design ändern möchten, klicken Sie auf „Seitenlayout" > „Designs", und wählen Sie „Facette".

b) Klicken Sie auf „Seitenlayout" > „Schriftarten", und wählen Sie „Office".

2) Ändern des Formats der Schriftart für Hyperlinks

a) Klicken Sie auf die Zelle mit dem Hyperlink. Klicken Sie auf der Registerkarte **Start** mit der rechten Maustaste auf die Formatvorlage **Hyperlink**, und wählen Sie **Ändern**.

b) Klicken Sie im Feld **Formatvorlage** auf **Format**.

c) Klicken Sie auf **Schriftart**, wählen Sie "Arial Black", und klicken Sie auf **OK**.

d) Klicken Sie auf **OK**, um das Feld **Formatvorlage** zu schließen.

3) Formatieren von Text in Kopf- oder Fußzeilen

a) Klicken Sie auf der Statusleiste auf die Schaltfläche **Seitenlayoutansicht**. 

b) Wählen Sie den Kopfzeilentext aus.

c) Wählen Sie auf der Registerkarte **Start** in der Gruppe **Schriftart** die Schriftart „Arial Black“ aus.

d) Wenn Sie fertig sind, klicken Sie auf der Statusleiste auf die Ansichtsschaltfläche **Normal**. 

4) Hinzufügen von Kommentaren

a. Klicken Sie auf die Zelle A1 und führen Sie eine der folgenden Aktionen aus:

1. Klicken Sie auf der Registerkarte Überprüfen in der Gruppe Kommentare auf Neuer Kommentar.
2. Klicken Sie mit der rechten Maustaste auf die Zelle, und klicken Sie dann auf **Kommentar einfügen**.
3. Es wird ein neuer Kommentar erstellt und der Zeiger zum Kommentar bewegt. In der Ecke der Zelle wird ein Indikator angezeigt.

b) Geben Sie im Nachrichtenteil des Kommentars den Kommentartext ein “**ok**”.

c) Klicken Sie außerhalb des Kommentarfelds.

d) Das Kommentarfeld wird nicht mehr angezeigt, aber der Kommentarindikator bleibt erhalten.

5) Anwenden bedingter Formatierungen mit Farbe in Excel

a) Wählen Sie die Spalte A, in der Sie die doppelten Werte mit einer Farbe formatieren möchten.

b) Klicken Sie auf der Registerkarte **Start** auf **Bedingte Formatierung** > **Regeln zum Hervorheben von Zellen > Textinhalt...**

c) Geben Sie „**PUBLIC HOLIDAY**“ ein.

d) Wählen Sie **Grüne Füllung mit dunkelgrünem Text**.

e) Klicken Sie auf **OK**, um Zellen zu formatieren.

f) Die doppelten Werte werden durch eine hellgrüne Füllung und mit dunkelgrünem Text hervorgehoben.

Now, open the tab called “Calculate” and perform tasks 6-8

6) Einfügen eines sich entfaltenden Kreisdiagramms

1. Wählen Sie in der Kalkulationstabelle die Daten aus, die Sie für das Kreisdiagramm verwenden möchten (A3 bis B8).
2. Klicken Sie auf **Einfügen** > **Kreis einfügen**.
3. Wählen Sie unter **2D-Kreisdiagramm** die am weitesten links stehende Option **Kreis**.
4. Um das Kreisdiagramm zu entfalten, führen Sie folgende Schritte aus:
   a) Klicken Sie auf das Diagramm, und wählen Sie dann den gesamten Kreis aus.
   b) Klicken Sie unter **Diagrammtools** auf die Registerkarte **Format**, und klicken Sie dann auf **Formatauswahl**.
   c) Ändern Sie im Bereich **Datenreihen formatieren** den Wert für den Prozentsatz

   im Feld **Kreisexplosion** unter **Option Reihe**, um den Kreis zu entfalten.

   Legen Sie den Wert auf 30 % fest.
5. Wenn Sie einen Titel zu Ihrem Diagramm hinzufügen möchten, wählen Sie das Diagramm aus, wählen Sie die Schaltfläche **Diagrammelemente** aus, und aktivieren Sie dann das Feld **Diagrammtitel**.
6. Wenn bereits ein , z. B. „Diagrammtitel“ vorhanden ist, ersetzen Sie ihn, indem Sie in das Titelfeld „Datum“ eingeben.
7. Klicken Sie unter **Diagrammtools** auf **Design** > **Diagrammelement hinzufügen**.
8. Zeigen Sie auf **Datenbeschriftungen**, und klicken Sie auf **Weitere Datenbeschriftungsoptionen**.
9. Klicken Sie im Bereich **Datenbeschriftungen formatieren** rechts auf **Beschriftungsoptionen**.
10. Deaktivieren Sie das Kontrollkästchen **Wert** ein, und überprüfen Sie das Feld **Prozentsatz**.
11. Klicken Sie auf  , klicken Sie auf **Füllung**, und wählen Sie die Schaltfläche **Farbverlauf aus.**

7) Einfügen eines Balken-aus-Kreisdiagramms

a) Wählen Sie in der Kalkulationstabelle das Kreisdiagramm aus.

b) Klicken Sie auf **Einfügen** > **Kreis einfügen**.

c) Wählen Sie unter **2D-Kreisdiagramm** die äußerste rechte Option **Balken aus Kreis**.
d) Zum Ändern der Farben, die im Diagramm verwendet werden, klicken Sie auf die Schaltfläche **Diagrammformatvorlagen** und dann auf **Farbe**.
e) Wählen Sie unter **Farbig** oder **Monochrom** ein Farbdesign aus, z. B. Farbe 4

8) Ausblenden von Kommentaren und ihren Indikatoren

a) Klicken Sie auf die Registerkarte Datei und dann auf Optionen.
b) Führen Sie in der Kategorie Erweitert unter Anzeige eine der folgenden Aktionen aus:

i. Damit sowohl Kommentare als auch Indikatoren in der gesamten Arbeitsmappe ausgeblendet werden, klicken Sie unter Für Zellen mit Kommentaren Folgendes anzeigen auf Keine Kommentare und Indikatoren.

Simplified Chinese:

MT

Open the first tab in the Excel spread sheet called “Calendar” and perform tasks from 1 to 5

1）快速更改工作表中的颜色、字体和效果

c）要切换到另一个主题，请单击页面布局>主题″，然后选择″方方面面″。

d）单击″页面布局″>字体，并选择″ Office ″。

2）更改超链接的字体 格式

e）单击包含超链接的单元 格 。在″主页″选项卡上右键单击该超链接 样式 和选择修改。

f）在“样式”对话框中，单击“格式”。

g）单击″字体″，选择″ Arial 黑色″，然后单击″确定″。

h）单击“确定”关闭“样式”对话框。

3）页眉或页脚中的文本设置 格式

e）单击状态 栏 上的"页面布局"视图"按钮。 

f）选择页眉文本。

g）在"开始"选项卡上的"字体"组中，选择" Arial 黑色"。

h）当您完成，请单击"普通"状态 栏 上的"视图"按钮。 

4）添加批注

a) 选择要向其添加批注的单元格，并执行下列操作之一：

1. 请在“审阅”选项卡的“批注”组中，单击“新建批注”。
2. 右键单击单元格，然后单击“插入批注”。

b) 一条新批注随即创建，指针会移到批注中。单元格的边角上会出现一个标记。

c) 在批注正文中，键入批注文字。

d) 在批注框外部单击。

e) 批注框消失，但批注标记仍然保留

5）在 Excel 中使用颜色应用条件格式

a）选择要使用颜色设置重复值的格式的列。

b）在“开始”选项卡上，单击“条件格式”>“突出显示单元格规则”>“文本包含”。

c）键入 "PUBLIC HOLIDAY"

d）选择 "绿填充色深绿色文本"

e）单击“确定”以设置单元格的格式。

f）重复值将使用浅红色填充和深红色文本突出显示。

Now, open the tab called “Calculate” and perform tasks 6-8

6）插入一个 exploding 饼图

1. 在您的电子表格中，选择数据以在饼图中使用。
2. 单击“插入”>“插入空间”。
3. 在"二维饼图"下，选择最左侧的选项，饼图"。

4. explode 饼图，请执行下列操作：

a) 单击图表，再选择整个饼图。

b. "图表工具"下，单击"格式"选项卡，然后单击"设置所选内容 格式"。

c. 在"设置数据系列，更改百分比窗 格 值

在"饼图分离程度"框的" 系列选项"下， explode 饼图。

将其设置为30% 。

5. 向图表中 添加 标题 ，请选择该图表，请选择"图表 元素 "按钮，然后选中"图表 标题 "框。

6. 如果存在已经 标题 ，如"图表 标题 "，通过键入将其替换 "日期"，在" 标题 "框中

7. 在"图表工具"下，单击"设计>添加 图表 元素"。

8. 指向"数据 标签"，然后单击"更多数据 标签 选项"。

9. 在""设置数据 标签 格式"右侧窗 格 中，单击" 标签 选项。

10. 取消选中"值"框并检查"百分比"框。

11. 单击  ，单击"填充"，然后选择"渐变填充"按钮。

7）插入一个复合条饼图

a) 在您的电子 表格 中，选择您的饼图。

b) 单击“插入”>“插入空间”。

c) 在"二维饼图"下，选择最右边的选项，"复合条饼图。

d) 若要更改图表中使用的颜色，请单击"图表 样式"按钮， ，单击"颜色"。

e) 在"选择颜色主题彩色或单色显示器免遭，如颜色4"。

8）显示或隐藏批注及其标记

a) 单击“文件”选项卡，然后单击“选项”。

b) 在“高级”类别的“显示”下，执行下列操作之一：

i. 若要隐藏整个工作簿中的批注和标记，请在“对于带批注的单元格，显示:”下单击“无批注或标识符”。

PE

Open the first tab in the Excel spread sheet called “Calendar” and perform tasks from 1 to 5

1）快速更改工作表中的颜色、字体和效果

a）要切换到另一个主题，请单击“页面布局”>“主题”，然后选择“方面”。

b）单击“页面布局”>“字体”，然后挑选“Office”。

2）更改超链接的字体格式

a）单击包含超链接的单元格。在“**开始**”选项卡上，右键单击“**超链接**”样式，然后挑选**修改**。

b）在“**样式**”对话框中，单击“**格式**”。

c）单击“**字体**”，选择“Arial Black”，然后单击“**确定**”。

d）单击“**确定**”关闭“**样式**”对话框。

3）设置页眉或页脚中文本的格式

a）在状态栏上，单击“**页面布局**”视图按钮。 

b）选择页眉文本。

c）在“**开始**”选项卡上的“**字体**”组中，选择“Arial Black”。

d）你完成后，单击状态栏上的“**普通**”视图按钮。 

4）添加批注

a) 选择要向其添加批注的单元格，并执行下列操作之一：

1. 请在“审阅”选项卡的“批注”组中，单击“新建批注”。
2. 右键单击单元格，然后单击“插入批注”。

b) 一条新批注随即创建，指针会移到批注中。单元格的边角上会出现一个标记。

c) 在批注正文中，键入批注文字。

d) 在批注框外部单击。

e) 批注框消失，但批注标记仍然保留

5）在 Excel 中使用颜色应用条件格式

a）挑选要使用颜色设置重复值的格式的列

b）在“**开始**”选项卡上，单击“**条件格式**”>“**突出显示单元格规则**”>“**文本包含...**”

c）键入“PUBLIC HOLIDAY”

d）选择“**绿填充色深绿色文本**”

e）单击“**确定**”以设置单元格的格式。

f）重复值将使用浅绿色填充和深绿色文本突出显示。

Now, open the tab called “Calculate” and perform tasks 6-8

6）插入一个分离型饼图

1. 在你的电子表格中，选择要在饼图中使用的数据（A3 到 B8）。
2. 单击“**插入**”>“**插入饼图**”。
3. 在“**二维饼图**”下，选择最左侧的“**饼图**”选项。
4. 要分离饼图，请执行下列操作：
   a) 单击图表，再选择整个饼图。
   b) 在“**图表工具**”下，单击“**格式**”选项卡，然后单击“**设置所选内容格式**”。
   c) 在“**设置数据系列格式**”窗格中，更改“**饼图分离程度**”框中“**系列选项**下的百分比值，以分离饼图。将其设置为 30% 。
5. 要添加图表标题，选择该图表，然后挑选“**图表元素**”按钮，再选中“**图表标题**”框。
6. 如果已经存在标题，如“图表标题”，请通过在标题框中键入“日期”替换该标题
7. 在“**图表工具**”下，单击“**设计**”>“**添加图表元素**”。
8. 指向“**数据标签**”，然后单击“**更多数据标签选项**”。
9. 在右侧的“**设置数据标签格式**”窗格中，单击“**标签选项**”。
10. 取消选中“**值**”框并选中“**百分比**”框。
11. 以此单击  、“**填充**”，然后选择“**渐变填充**”按钮。

7）插入一个复合条饼图

a）在你的电子表格中，选择你的饼图。

b）单击“**插入**”>“**插入饼图**”。

c）在“**二维饼图**”下，选择最右边的“**复合条饼图**”选项。

d）若要更改图表使用的颜色，请单击“**图表样式**”按钮，然后单击“**颜色**”。

e）在“**彩色**”或“**单色**”下，选择颜色主题，如“颜色 4”。

8）**显示或隐藏批注及其标记**

a) 单击“文件”选项卡，然后单击“选项”。

b) 在“高级”类别的“显示”下，执行下列操作之一：

i. 若要隐藏整个工作簿中的批注和标记，请在“对于带批注的单元格，显示:”下单击“无批注或标识符”。

Japanese:

MT

Open the first tab in the Excel spread sheet called “Calendar” and perform tasks from 1 to 5

1) 色、フォント、および効果、ワークシート内をすばやく変更します。

a) 別のテーマに切り替えるに、[ページ レイアウト] をクリックして > テーマ、および「ファセット」を選択します。

b) [ページ レイアウト] をクリックして > フォント、および"Office"を選択します。

2) ハイパーリンクのフォントの書式設定を変更します。

a) ハイパーリンクを含むセルをクリックします。[ホーム] タブの [ハイパーリンクのスタイルを右クリックし、変更を選びます。

b) スタイル] ボックスで、[書式設定] をクリックします。

c) [フォント] をクリックし、「明朝」[ok] をクリックします。

d) [スタイル] ボックスを閉じるには、[ok] をクリックします。

3) ヘッダーまたはフッターのテキストの書式設定

a) ステータス バーで、[ページ レイアウト ビュー] ボタンをクリックします。

b) ヘッダーのテキストを選択します。

c) [ホーム] タブの [フォント] グループで「明朝」を選びます。

d) 完了したら、ステータス バーの [標準表示モード] ボタンをクリックします。

4) コメントを追加する

a) コメントを追加するには、セルA1を選択し、次のいずれかの操作を行います。

1. [校閲] タブの [コメント] グループで [コメントの挿入] をクリックします。
2. セルを右クリックし、[コメントの挿入] をクリックします。
3. 新しいコメントが作成され、カーソルがコメントに移動します。セルの隅にはインジケーターが表示されます。

b) コメントの本文に、「OK」を入力します。

c) コメント ボックスの外側をクリックします。

d) コメント ボックスは消えますが、コメント インジケーターは残ります。コメントが表示されたままにするには、次の操作を行います。

5) Excel での色では、条件付き書式を適用します。

a) 列 A の選択を色で重複する値の書式を設定します。

b) [ホーム] タブの [条件付き書式 > セルの強調表示ルール > が含まれるテキスト.

c) 種類「PUBLIC HOLIDAY」

d) 緑の塗りつぶしと濃い緑のテキストを選択します。

e) セルの書式を設定するには、[ok] をクリックします。

f) 重複する値が明るい緑色の塗りつぶしと濃い緑のテキストを強調表示されます。

Now, open the tab called “Calculate” and perform tasks 6-8

6) 分解を円グラフを挿入します。

1. スプレッドシートで、円グラフに使用するデータを選択します (B8 に A3)。
2. [挿入] をクリックして > 円を挿入します。
3. 2-d 円グラフ] の下には、左端 [円] オプションを選択します。
4. 円グラフを分割するのには、次の手順で行います。
   a) 図をクリックし、[円グラフ全体を選択します。
   b) [グラフ ツール] の [書式] タブをクリックし、[書式の選択] をクリックします。
   c) データ系列の書式設定] ウィンドウで、[パーセンテージ] の値を変更します。

   [円グラフの切り離し] ボックスの [系列] オプションの円グラフを分割する] の [します。 30% に設定します。
5. グラフにタイトルを追加するには、グラフを選択し、グラフの要素を選びます

   このボタンをクリックすると、し、[グラフ　タイトル]　ボックスを確認します。
6. 既にある場合、「グラフ タイトル」などのタイトルを置き換えることを入力 [タイトル] ボックスには、「日付」
7. [グラフ ツール] の [デザイン] をクリックして > グラフの要素を追加します。
8. データ ラベル] をポイントし、[その他のデータ ラベルのオプション] をクリックします。
9. ウィンドウで、データ ラベルの書式設定、右側に [ラベル オプション] をクリックします。
10. [値] ボックスをオフにし、パーセンテージ ボックスをオンにします。
11.  クリックし、塗りつぶし、グラデーションの塗りつぶし] ボタンを選びます。

7) 円グラフのバーを挿入します。

a) スプレッドシートで、円グラフを選択します。

b) [挿入] をクリックして > 円を挿入します。

c) [2-d 円グラフ] の [バーの円グラフ、右端のオプションを選択します。

d) グラフを使用する色を変更するには、[グラフ スタイル] ボタンをクリックし、[色] をクリックします。

e) [カラフル] または [Monochromatic、4 の色などのテーマの色を選びます。

8) コメントおよびコメントインジケーターの非表示

a) [ファイル] タブをクリックし、[オプション] をクリックします。

b) [詳細設定] カテゴリの [表示] で、次のいずれかの操作を行います。

i. ブック全体でコメントとインジケーターの両方を非表示にするには、[コメントのあるセルに対して表示] の [コメントとインジケーター両方なし] をクリックします。

PE

Open the first tab in the Excel spread sheet called “Calendar” and perform tasks from 1 to 5

1) ワークシートの色、フォント、および効果をすばやく変更します。

a) 別のテーマに切り替えるには、[ページ レイアウト] タブの [テーマ] をクリックし、[ファセット] を選択します。

b) [ページ レイアウト] タブの [フォント] をクリックし、[Office] を選択します。

2) ハイパーリンクのフォントの書式設定を変更します。

a) ハイパーリンクを含むセルをクリックします。[ホーム] タブの [ハイパーリンク] のスタイル] を右クリックし、[変更] を選択します。

b) [スタイル] ボックスで、[書式設定] をクリックします。

c) [フォント] をクリックし、「明朝」を選択し、[OK] をクリックします。

d) [スタイル] ボックスを閉じるには、[OK] をクリックします。

3) ヘッダーまたはフッターのテキストの書式設定

a) ステータス バーの [ページ レイアウト ビュー] ボタンをクリックします。

b) ヘッダーのテキストを選択します。

c) [ホーム] タブの [フォント] グループから「明朝」を選択します。

d) 完了したら、ステータス バーの [標準] 表示ボタンをクリックします。 

4) コメントを追加する

a) コメントを追加するには、セルA1を選択し、次のいずれかの操作を行います。

1. [校閲] タブの [コメント] グループで [コメントの挿入] をクリックします。
2. セルを右クリックし、[コメントの挿入] をクリックします。
3. 新しいコメントが作成され、カーソルがコメントに移動します。セルの隅にはインジケーターが表示されます。

b) コメントの本文に、「OK」を入力します。

c) コメント ボックスの外側をクリックします。

d) コメント ボックスは消えますが、コメント インジケーターは残ります。コメントが表示されたままにするには、次の操作を行います。

5) Excel では、条件付き書式を適用して色を変更します。

a) 重複する値に色を付ける場合、列 A を選択します。

b) [ホーム] タブの [条件付き書式] で、[セルの強調表示ルール] をポイントし、[次を含むテキスト] をクリックします。

c) 「PUBLIC HOLIDAY」を入力します。

d) [濃い緑の文字、緑の背景] を選択します。

e) [OK] をクリックし、セルの書式を設定します。

f) 重複した値は濃い緑の文字、明るい緑の背景で強調表示されます。

6) 分割した円グラフを挿入します。

1. スプレッドシートで、円グラフに使うデータを選択します (A3 から B8)。
2. [挿入] タブの [円の挿入] をクリックします。
3. [2-D 円] の左端の [円] を選択します。
4. 円グラフを展開するには、次の手順で行います。

a) 図をクリックし、円グラフ全体を選択します。
b) [グラフ ツール] の [書式] タブをクリックし、[選択範囲のフォーマット] をクリックします。
c) [データ系列の書式設定] ウィンドウの [系列のオプション] の [円グラフの切り離し] ボックスで、 パーセント値を変更し、円グラフを分割させます。

30% に設定します。

5. グラフにタイトルを追加するには、グラフを選択し、[グラフ要素] ボタンを選択し、 [グラフ タイトル] ボックスをチェックします。
6. タイトルが既にある場合、タイトル ボックスに「日付」を入力し、「グラフ タイトル」などのタイトルに置き換えます。
7. [グラフ ツール] の [デザイン] タブの [グラフ要素を追加] をクリックします。
8. [データ ラベル] をポイントし、[その他のデータ ラベル オプション] をクリックします。
9. ウィンドウの右側に現れた [データ ラベルの書式設定] ウィンドウで、[ラベル オプション] をクリックします。
10. [値] ボックスをオフにし、[パーセント] ボックスをオンにします。
11.  をクリックし、[塗りつぶし] タブで [塗りつぶし (グラデーション)] チェック ボックスをオンにします。

7) 補助縦棒付き円グラフを挿入します。

a) スプレッドシートで、円グラフを選択します。
b) [挿入] タブの [円の挿入] をクリックします。
c) [2-D 円] の右端の [補助縦棒付き円] を選択します。
d) グラフに使用する色を変更するには、[グラフ スタイル] ボタンをクリックし、[色] をクリックします。
e) [カラフル] または [モノクロ]、色 4 などのテーマの色を選択します。

8) コメントおよびコメントインジケーターの非表示

a) [ファイル] タブをクリックし、[オプション] をクリックします。
b) [詳細設定] カテゴリの [表示] で、次のいずれかの操作を行います。

i. ブック全体でコメントとインジケーターの両方を非表示にするには、[コメントのあるセルに対して表示] の [コメントとインジケーター両方なし] をクリックします。

APPENDIX C: Baseline

English

Microsoft Office

An office suite is a collection of bundled productivity software intended to be used by knowledge workers. The components are generally distributed together, have a consistent user interface and usually can interact with each other, sometimes in ways that the operating system would not normally allow. Existing office suites contain wide range of various components. Most typically, the base components include, Word processors, Spreadsheets, Presentation programs.

Microsoft Office is an office suite of desktop applications, servers and services for Microsoft Windows and OS X operating systems. Initially, the first version of Office contained Microsoft Word, Microsoft Excel and Microsoft PowerPoint. Over the years, Office applications have grown substantially closer with shared features such as a common spell checker.

Microsoft supports Office for the Windows and OS X platforms, as well as mobile versions for Windows Phone, Android and iOS platforms. Microsoft has stated that it plans to create a version of Office for "other popular platforms" as well.

Press **F10** to start

German

Microsoft Office

Ein Office-Paket ist eine Sammlung gebündelter Produktivitätssoftware die für die Nutzung von Wissensarbeitern bestimmt ist. Die Komponenten werden in der Regel zusammen veröffentlicht, haben eine einheitliche Benutzeroberfläche und können meist miteinander interagieren, jedoch manchmal auf eine Art, die das Betriebssystem normalerweise nicht unterstützen würde. Bestehende Office-Pakete beinhalten eine große Auswahl verschiedener Komponenten. Die Basiskomponenten beinhalten meistens Textverarbeitungsprogramme, Tabellenkalkulationen und Präsentationsprogramme.

Microsoft Office ist ein Office-Paket, welches aus Desktop-Anwendungen, Servern und Diensten für die Betriebssysteme Microsoft Windows und OS X besteht.

Ursprünglich enthielt die erste Office-Version Microsoft Word, Microsoft Excel und Microsoft PowerPoint. Im Laufe der Jahre sind Office-Anwendungen durch gemeinsam genutzte Funktionen, wie z.B. eine gemeinsame Rechtschreibprüfung, wesentlich näher zusammen gewachsen.

Microsoft unterstützt Office sowohl auf Windows oder OS X Plattformen, als auch auf mobilen Versionen für das Windows Phone, auf Android und iOS Plattformen. Microsoft hat angegeben, dass sie außerdem eine Office-Version für “andere gängige Plattformen” erstellen wollen.

Press F10 to start

Simplified Chinese

Microsoft Office

OFFICE套组是一组给知识工作者使用的创作软件套装。组件通常都是一起贩售，有相同的使用者介面且通常可以相互运行，有时候连 在一般作业系统不允许的情况下也可相互运行。现有的OFFICE套组的广大系列中含有多种不同的组件。最具代表性的基础组件包含，文书处理器、电子试算表、简报软体。

Microsoft Office是办公室使用的桌面应用程式套组，为Microsoft Windows 和 OS X 作业所设计。最初，第一版的OFFICE包含了Microsoft Word, Microsoft Excel 和 Microsoft PowerPoint。几年下来，OFFICE应用程式大幅地成长，特别是可共用的功能，例如通用的拼字检查。

Microfost OFFICE支援WINDOWS和OS X作业平台，也支援手机版本的Windows Phone, Android 和 iOS平台。Microsoft声明也计画为”其他热门的平台”制作可支援的OFFICE版本。

Press F10 to start

Japanese

Microsoft Office

オフィススイートとは、知識労働者向けの複数のプロダクティビティ・ソフトウェアを、ひとつにまとめたソフトウェアのことである。一般的にはオフィススイートを構成する個々のアプリケーションは、一揃いのパッケージとして提供されていたり、ユーザーインターフェースが統一されていたり、通常にアプリケーション間でデータのやりとりができる。そのやりとりで通常のシステム許可が必要ない場合もある。入手できるオフィススイートはさまざまなアプリケーションで構成されており、通常にワープロソフトや表計算ソフトやプレゼンテーションソフトなどの基本となるアプリケーションが含まれている。

Microsoft　Officeは、デスクトップ製品、サーバー、サービスなどが含まれており、Microsoft WindowsやOS Xオペレーティング・システムで利用可能なオフィススイートである。最初に販売されたバージョンはMicrosoft　WordとMicrosoft Excel、Microsoft PowerPointが含まれたものであった。年月と共にOfficeのアプリケーションの類似性が高まっており、現在はスペル・チェッカーなどのような共通した機能が用意されている。

マイクロソフトはOfficeをWindows やOS Xのプラットフォームに対応しており、モバイル版をWindows PhoneやAndroid、iOSのプラットフォームにも対応している。「他に人気があるプラットフォーム」向けのバージョンも開発する予定だとマイクロソフトは発表した。

Press F10 to start

APPENDIX D: Visit Duration Results

Baseline

As seen in Figure 0:1 the German PEz group has longer fixations on the AOI baseline when compared to the other groups. However, no statistically significant differences were found for any of the groups compared. This lack of significantly differences indicates that in general, all groups had the same level of cognitive effort required when reading the text.

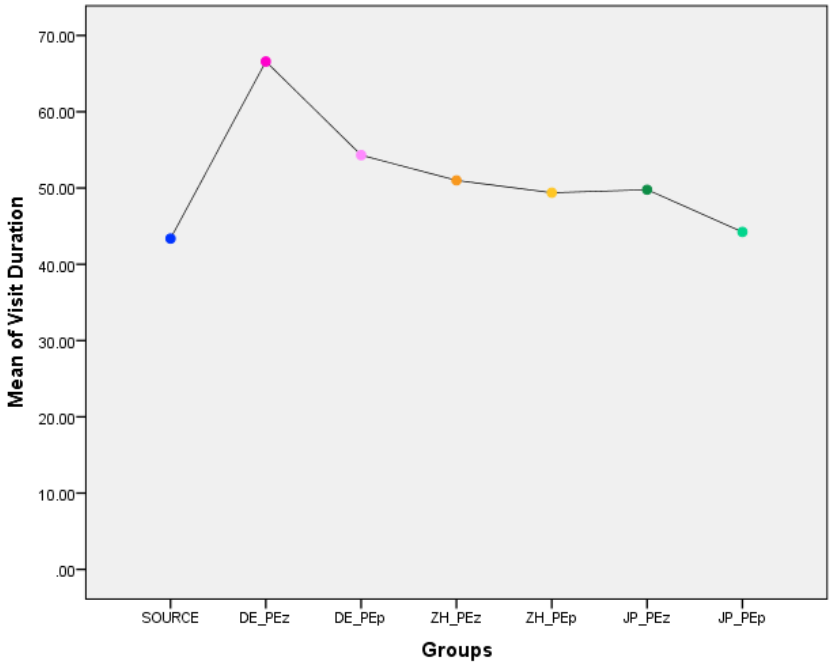


Figure 0:1 -Visit duration - Baseline

MT Instructions

A two-way MANOVA with repeated measures was conducted in order to compare whether the factors Language and PE_LEVEL have an effect on Visit Duration (VD) for both AOIs: Instruction (VD_INST) and User Interface (VD_UI).

LANGUAGE: The factor Language was found not to have a statistically significant difference on VD, where (F (2, 35) = .17, p>.10). This means that when the factor language is considered without distinctions between PE_LEVELs, there is no statistically significant differences across the three translated languages DE (M=529.70, SE=58.73), ZH (M=571.77, SE=54.63), JP (M=534.01, SE=58.73).

POST-EDITING LEVEL: The factor PE_LEVEL was also found not to have a statistically significant difference on VD, where (F (1, 35) = .36, p>.10). This means that when the factor PE_LEVEL is considered without distinctions between languages, there is no statistically significant differences across the two post-editing levels PEz (M=565.27, SE=46.07), and PEp (M=525.04, SE=47.64).

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect on VD, where (F (2, 35) = .77, p>.10). This means that the factor language combined with the factor PE_LEVEL do not have a joint effect on VD.

Table 0:1 shows the mean and standard deviation for each language and their respective post-editing levels for each AOI (instructions and user interface).

AOIs	Groups		Mean	Std. Deviation
VD_INST	DE	PEz	463.99	56.39
		PEp	452.29	99.74
	ZH	PEz	507.65	207.99
		PEp	362.13	106.97
	JP	PEz	427.06	201.50
		PEp	365.90	166.00
VD_UI	DE	PEz	628.12	146.56
		PEp	574.40	232.10
	ZH	PEz	779.11	361.03
		PEp	638.19	321.99
	JP	PEz	585.71	254.82
		PEp	757.37	358.83

Table 0:1 - Mean and Standard Deviation for Visit Duration - Translated Content

The test of within-subjects determined that VD did not have any significant effects in the interactions with Language (F (2, 35) = 2.18, p>.10), PE_LEVEL (F (1, 35) = 1.23, p>.10) or Language*PE_LEVEL (F (2, 35) = 2.03, p>.10). However, there was a statistically significant difference between VD_INST and VD_UI (F (1, 35) = 62.05, p<.001). Figure 0:2 illustrates the estimated marginal means for each post-editing level for visit duration instructions, while Figure 0:3 illustrates for visit duration UI.

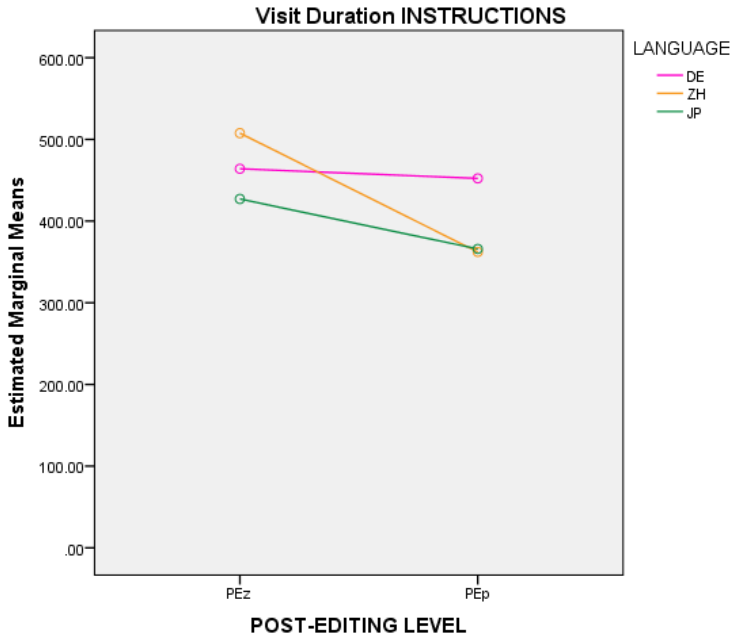


Figure 0:2-Visit Duration Instructions - Translated Content

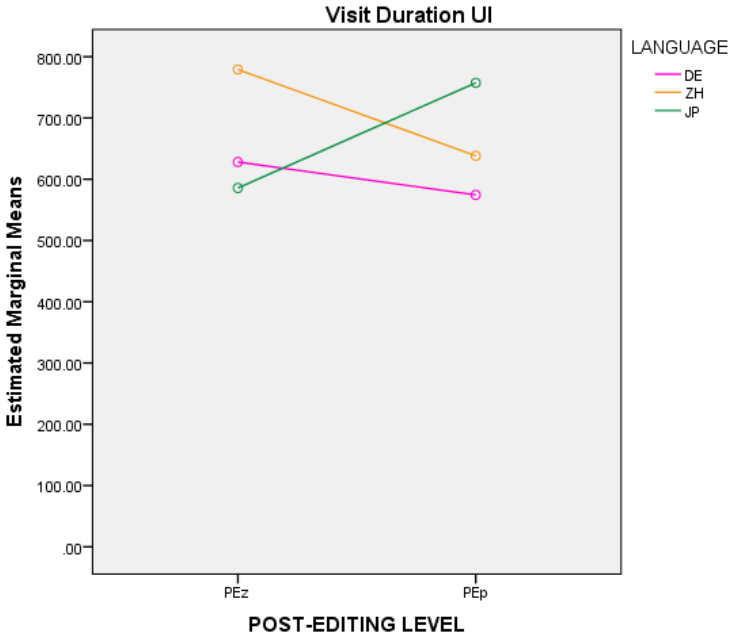


Figure 0:3 - Visit Duration UI - Translated Content

When looking at the AOI INST across groups, all PEp groups present shorter visit duration when compared to their PEz groups. However, these results are only statistically significant for the Simplified Chinese language (ZH_PEz (M= 507.65, SD=207.99), ZH_PEp (M=362.13, SD=106.97)) at the *p* <.10 level.

When looking at the AOI UI across groups, the JP_PEp group presents longer visit duration when compared to JP_PEz, which differs from all the German and Simplified

Chinese languages where the longer visit duration is seen in the PEz groups. However, these results were found not to be statistically significant.

Figure 0:4 shows the differences between VD_INST and VD_UI for each group. When comparing both AOIs (INST *vs* UI), all groups have longer visit duration in the AOI user interface when compared to the AOI instructions, which means that all groups spent more time in the user interface. These results were statistically significant for the groups DE_PEz, ZH_PEz, ZH_PEp, JP_PEz, and JP_PEp at the $p<.05$ level; apart from the DE_PEp group which was not statistically significant at the $p>.10$ level.

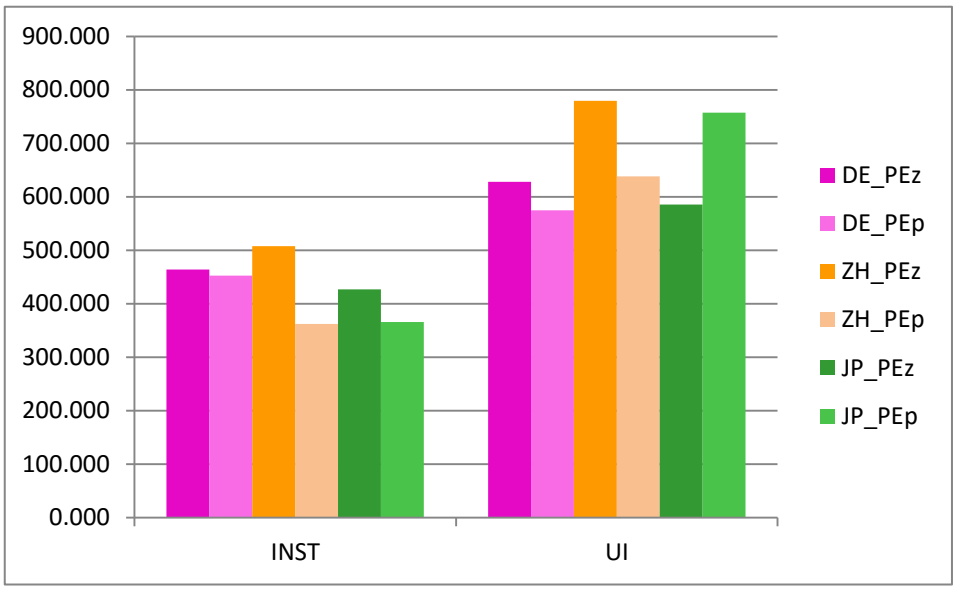


Figure 0:4- Differences per group for Visit Duration – Translated Content

Comparison with Source

The performance of the participants who used the English Source of the MT Instructions was also computed for visit duration via a one-way MANOVA with repeated measures. Table 0:2 shows the mean and standard deviation for each language and their respective post-editing levels (in seconds) for each AOI (instructions and UI) compared to the English Source.

AOIs	Groups		Mean	Std. Deviation
VD_INST	EN	SOURCE	293.12	37.72
	DE	PEz	463.99	56.39
		PEp	452.29	99.74
	ZH	PEz	507.65	207.99
		PEp	362.13	106.97
	JP	PEz	427.06	201.50
		PEp	365.90	166.00
VD_UI	EN	SOURCE	442.51	94.885
	DE	PEz	628.12	146.56
		PEp	574.40	232.10
	ZH	PEz	779.11	361.03
		PEp	638.19	321.99
	JP	PEz	585.71	254.82
		PEp	757.37	358.83

Table 0:2 - Mean and Standard Deviation for Visit Duration (secs) - Source

The factor PE_LEVEL was found not to have a statistically significant effect on VD (F (6, 42) = 1.24, p>.10). The test of within-subjects determined that VD did not have any significant effects in the interaction with PE_LEVEL (F (6, 42) = 1.43, p>.10). However, there was a statistically significant difference between VD_INST and VD_UI (F (1, 42) = 72.22, p<.001). A pairwise comparison found that the participants who used the source instructions (EN (M=367.81, SE=73.63)) had shorter VD when compared to the DE_PEz group (M=546.05, SE=73.63), at the p <.10 level; ZH_PEz (M=643.38, SE=73.63), at the p<.05 level; and JP_PEp (M=545.39, SE=73.63), at the p<.10 level.

Figure 0:5 illustrates the estimated marginal means for each language and their PE_LEVEL compared to the Source for the AOI instructions, while Figure 0:6 shows the means for each language and their PE_LEVEL compared to the source for the AOI user interface.

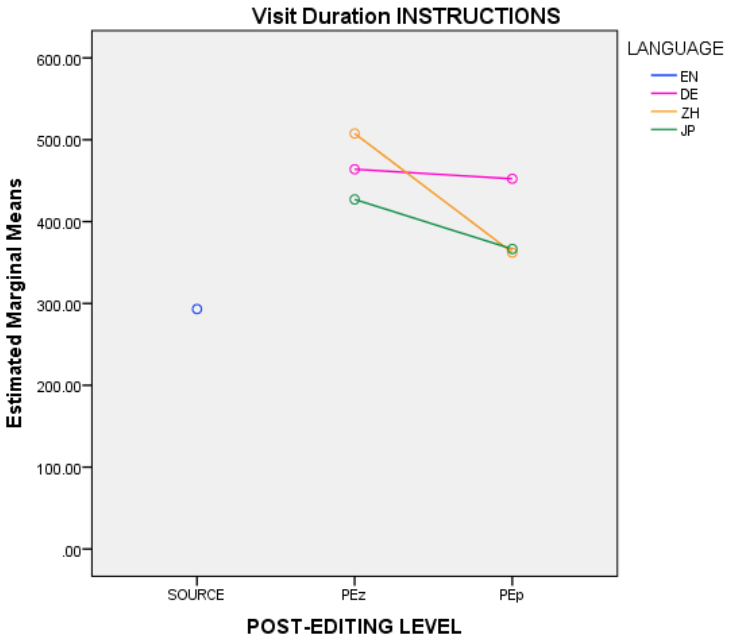


Figure 0:5 - Visit Duration Instructions (secs) - Source

When looking at the AOI INST across groups (Figure 0:5), the EN_Source group presents shorter visits in the instruction when compared to all the groups. However, these effect was statistically significant only against the DE_PEz, DE_PEp (p<.05), ZH_PEz (p<.005), and JP_PEz (p<.10) groups.

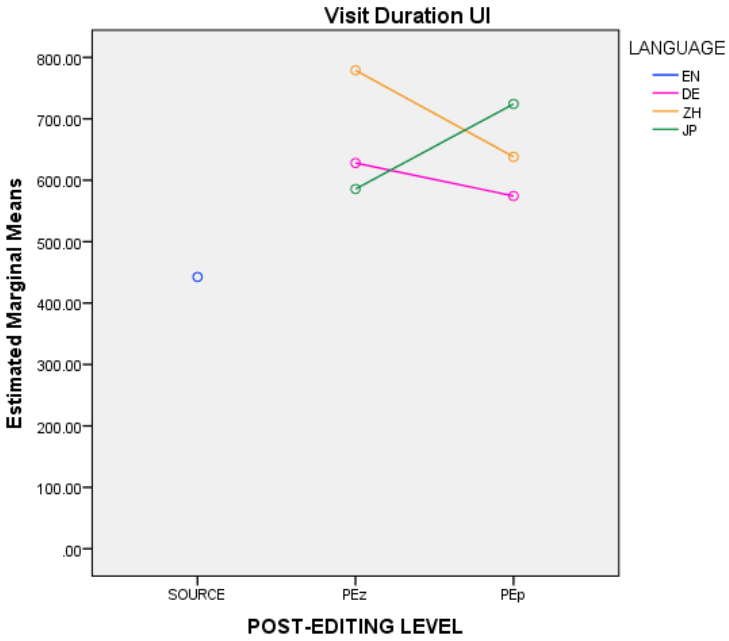


Figure 0:6 - Visit Duration UI (secs) - Source

When looking at the AOI UI across groups (Figure 0:6), the EN_Source group presents shorter visits in the UI when compared to all the groups. However, this effect

was statistically significant only against the ZH_PEz and JP_PEp groups at the p<.05 level.

Figure 0:7 shows the differences between VD_INST and VD_UI for each group compared to the EN_Source. The source also presents higher visit duration in the AOI UI when compared to the AOI INST. This result was statistically significant for the EN_Source group at the p <.05 level. This means that for all groups, including for the EN_Source group, there was more cognitive effort related to the user interface window against the instruction window.

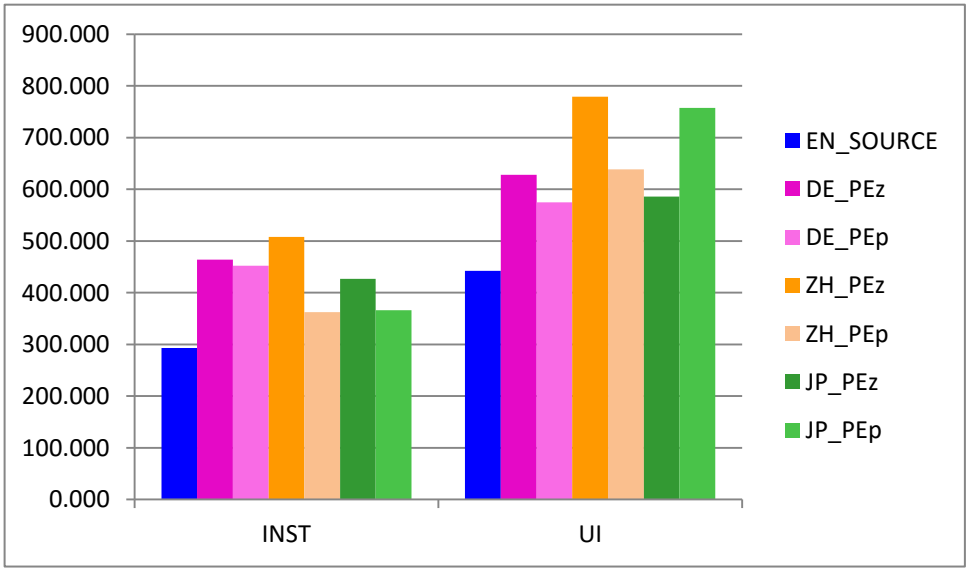


Figure 0:7 - Differences per group for Visit Duration – Source

HT Instructions

A two-way MANOVA with repeated measures was conducted in order to compare whether the factors Language and PE_LEVEL have an effect on Visit Duration (VD) for both AOIs: Instruction (VD_INST) and User Interface (VD_UI) for the HT Instructions.

LANGUAGE: The factor Language was found not to have a statistically significant difference on VD, where (F (2, 35) = 1.56, p>.10). This means that when the factor language is considered without distinctions between PE_LEVELs, there is no statistically significant differences across the three translated languages DE (M=91.49, SE=10.34), ZH (M=100.72, SE=9.62), JP (M=75.86, =10.34).

POST-EDITING LEVEL: The factor PE_LEVEL was also found not to have a statistically significant difference on VD, where (F (1, 35) = .63, p>.10). This means that when the factor PE_LEVEL is considered without distinctions between languages, there

is no statistically significant differences across the two post-editing levels PEz (M=92.16, SE=8.11), and PEp (M=86.55, SE=8.39).

INTERACTION: The interaction Language*PE_LEVEL was found not to have a statistically significant effect on VD, where (F (2, 35) = .05, p>.10). This means that the factor language combined with the factor PE_LEVEL do not have a joint effect on VD.

Table 0:3 shows the mean and standard deviation for each language and their respective post-editing levels per AOI (instructions and UI) for the HT instructions.

AOIs	Groups		Mean	Std. Deviation
VD_INST	DE	PEz	89.53	13.35
		PEp	96.88	28.22
	ZH	PEz	99.37	32.66
		PEp	78.80	13.85
	JP	PEz	72.56	35.60
		PEp	65.67	31.07
VD_UI	DE	PEz	101.91	27.33
		PEp	77.65	50.98
	ZH	PEz	110.31	54.29
		PEp	114.40	32.83
	JP	PEz	79.28	50.89
		PEp	85.94	80.43

Table 0:3 - Mean and Standard Deviation for Visit Duration (secs) - HT Instructions

The test of within-subjects determined that VD did not have a significant effect in the interactions with Language (F (2, 35) = 2.29, p>.10) or PE_LEVEL (F (1, 35) =0.45, p>.10). A significant effect was found for Language*PE_LEVEL (F (2, 35) = 2.74, p<.10). There was also a statistically significant difference between VD_INST and VD_UI (F (1, 35) = 4.61, p<.05).

Figure 0:8 illustrates the estimated marginal means for each post-editing level for visit duration instructions, while Figure 0:9 illustrates for fixation duration UI for the HT Instructions.

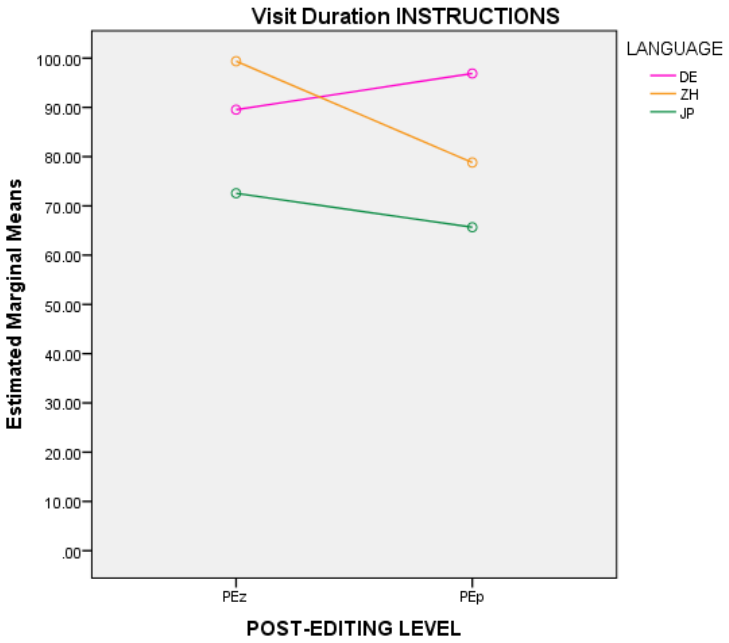


Figure 0:8 - Visit Duration Instructions (secs) - HT Instructions - Translated Content

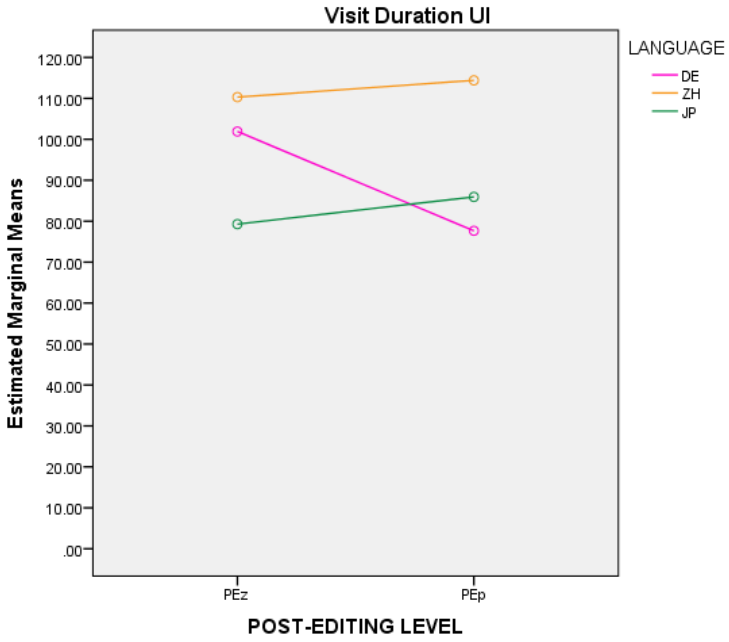


Figure 0:9 - Visit Duration UI (secs) - HT Instructions - Translated Content

When looking at the AOI INST across PE_LEVELs (Figure 0:8), longer visits can be observed for the ZH_PEz and JP_PEz groups when compared to their PEp groups, which indicates that the groups which used the raw machine translated version of the instructions had more cognitive effort observed when reading the instructions. The German language interestingly shows longer visits in the AOI INST group for the PEp group when compared to the PEz group. However, none of the results were statistically significant (p >.10).

When looking at the AOI UI across PE_LEVELS (Figure 0:9), longer visits can be observed for the ZH_PEp and JP_PEp groups when compared to their PEz groups, which indicates that the groups which used the post-edited version of the instructions had more cognitive effort observed when using the UI. The German language, again, oppositely from the other languages, shows longer visits for the PEz group. However, neither results were statistically significant (p >.10).

Figure 0:10 shows the differences between VD_INST and VD_UI for each language and post-editing level for the AOIs instructions and UI. Apart from the DE_PEp groups, all the other groups have longer visits in the AOI user interface when compared to the AOI instruction. These results, however, are statistically significant only for the ZH_PEp group at the p<.005 level

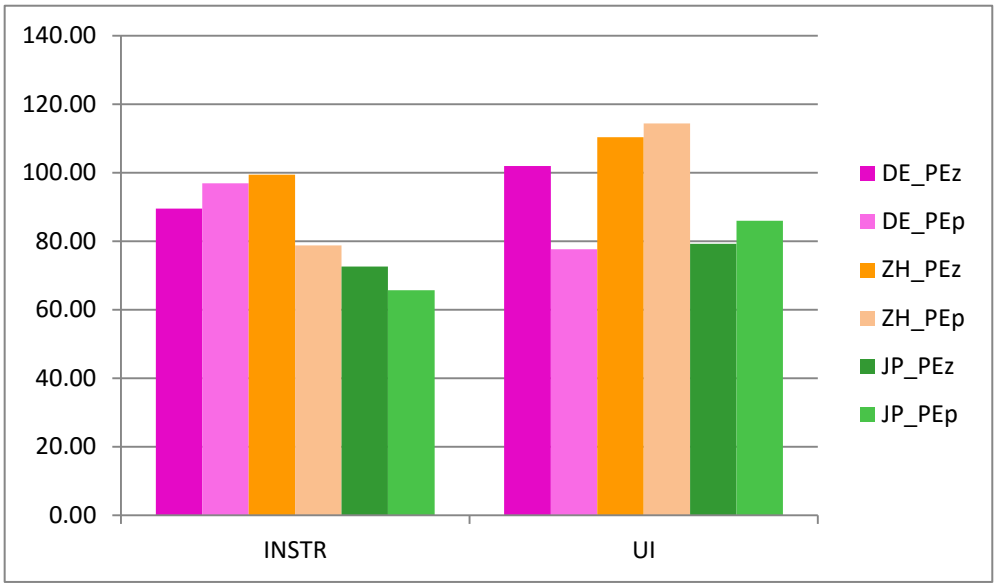


Figure 0:10 - Differences per group for Visit Duration (secs) - HT Instructions - Translated Content

Comparison with Source

The performance of the participants who used the English Source of the HT Instructions was also computed for visit duration via a one-way MANOVA with repeated measures. Table 0:4 shows the mean and standard deviation for each language and their respective post-editing levels (in seconds) for each AOI (instructions and UI) compared to the English Source.

AOIs	Groups		Mean	Std. Deviation
VD_INST	EN	SOURCE	57.06	16.45
	DE	PEz	89.53	13.35
		PEp	96.88	28.22
	ZH	PEz	99.37	32.66
		PEp	78.80	13.85
	JP	PEz	72.56	35.60
		PEp	65.67	31.07
VD_UI	EN	SOURCE	64.09	27.422
	DE	PEz	101.91	27.33
		PEp	77.65	50.98
	ZH	PEz	110.31	54.29
		PEp	114.40	32.83
	JP	PEz	79.28	50.89
		PEp	85.94	80.43

Table 0:4 - Mean and Standard Deviation for Visit Duration (secs) - HT Instructions - Source

The factor PE_LEVEL was found not to have a statistically significant effect on VD (F (6, 42) = 1.37, p>.10). The test of within-subjects determined that VD (when both AOIs are considered) had a significant effect in the interaction with PE_LEVEL (F (6, 42) = 1.91, p<.10). There was also a statistically significant difference between VD_INST and VD_UI (F (1, 42) = 5.28, p<.05). A pairwise comparison found that the participants who used the source instructions (EN (M=60.57, SE=13.19)) had shorter VD when compared to the DE_PEz (M=95.72, SE=13.19), ZH_PEz (M=, SE=) and ZH_PEp (M=96.60, SE=12.33), at the p<.10 level.

Figure 0:11 illustrates the estimated marginal means for each language and their PE_LEVEL compared to the Source for the AOI instructions, while Figure 0:12 shows the means for each language and their PE_LEVEL compared to the source for the AOI user interface.

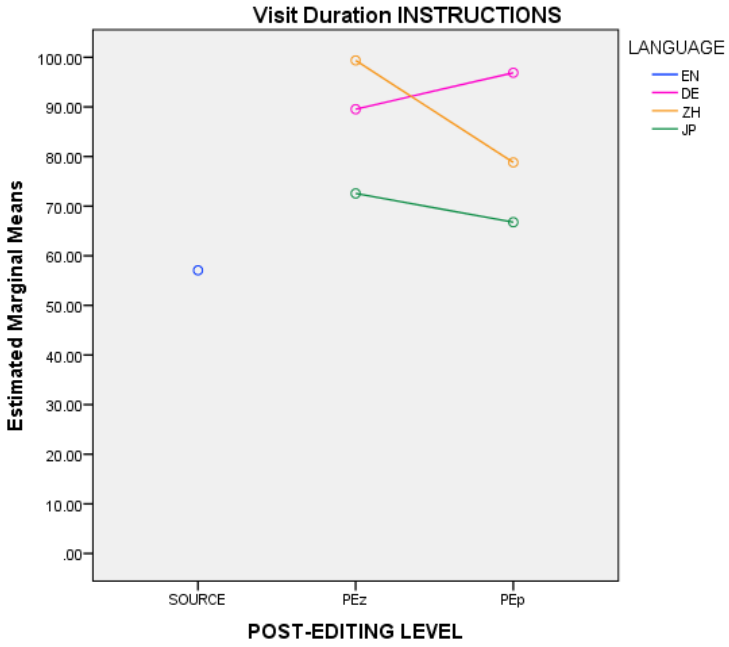


Figure 0:11 - Visit Duration Instructions (secs) - HT Instructions - Source

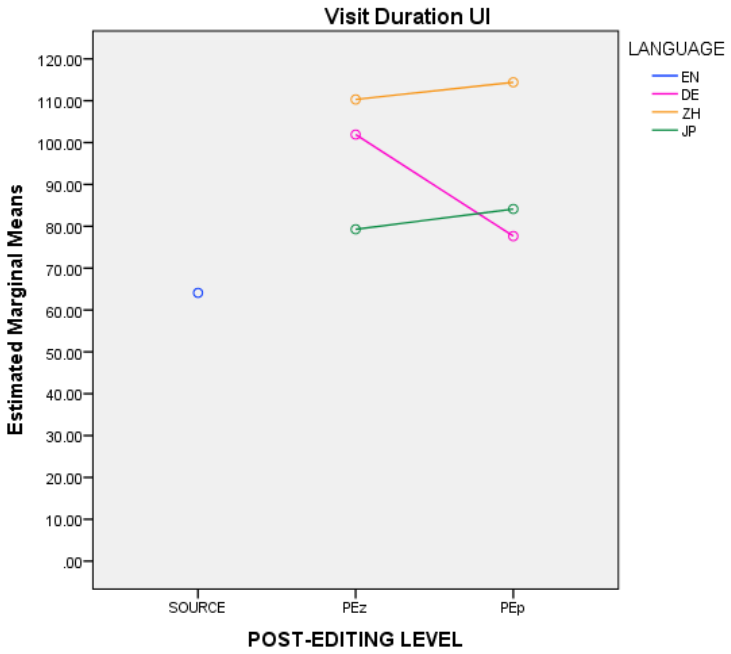


Figure 0:12 - Visit Duration UI - HT Instructions - Source

When looking at the AOI INST across groups (Figure 0:11), the EN_Source group presents shorter visits in the instruction when compared to all the groups. However, these effect was statistically significant only against the DE_PEz (p<.05), DE_PEp (p<.005), ZH_PEz (p<.005), and ZH_PEp (p<.10) groups.

When looking at the AOI UI across groups (Figure 0:12), the EN_Source group presents shorter visits in the UI when compared to all the groups. However, this effect

was statistically significant only against the ZH_PEz (p<.10) and ZH_PEp (p<.05 level) groups.

Figure 0:13 shows the differences between VD_INST and VD_UI for each group compared to the EN_Source. The source also follows the previous results in which the translated groups the groups have longer visits in the AOI user interface when compared to the AOI instruction (apart from the DE_PEp group). This result, however, was not statistically significant for the EN_Source group at the p >.10 level.

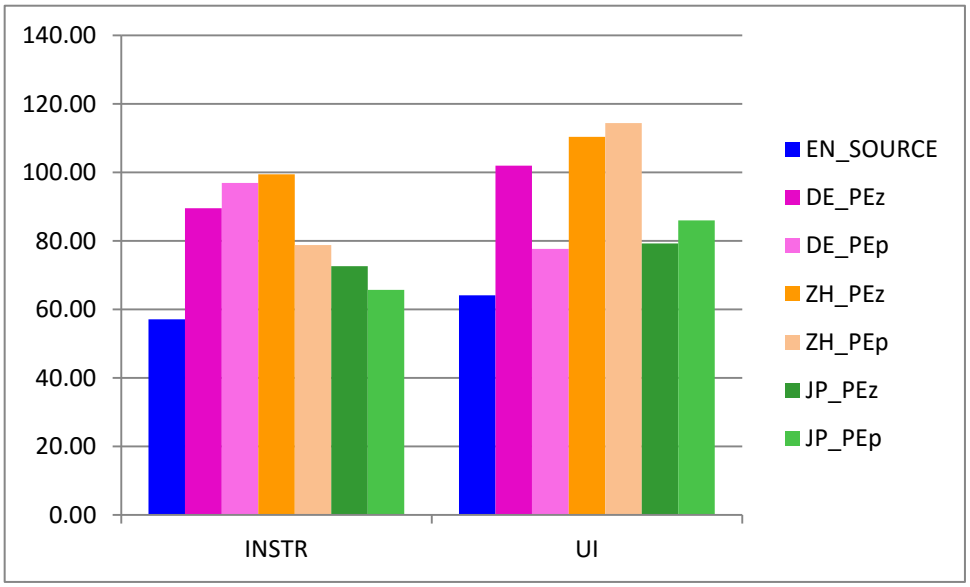


Figure 0:13 - Differences per group for Visit Duration - Source - HT Instructions