Post-Editing Evaluations: Trade-offs between Novice and Professional Participants

Joss Moorkens
ADAPT Centre/School of Computing
Dublin City University
Ireland
joss.moorkens@dcu.ie

Sharon O'Brien
ADAPT Centre/SALIS/CTTS
Dublin City University
Ireland
sharon.obrien@dcu.ie

Abstract

The increasing use of post-editing in localisation workflows has led to a great deal of research and development in the area, much of it requiring user evaluation. This paper compares some results from a post-editing user interface study carried out using novice and expert translator groups. By comparing rates of productivity, edit distance, engagement with the research, and qualitative findings regarding each group's attitude to post-editing, we find that there are trade-offs to be considered when selecting participants for evaluation tasks. Novices may generally be more positive and enthusiastic and will engage considerably with the research while professionals will be more efficient, but their routines and attitudes may prevent full engagement with research objectives.

1 Introduction

The use of machine translation (MT) in commercial translation and localisation workflows has grown exponentially in recent years. Relatively recent breakthroughs in the quality of statistical machine translation (SMT) output has led to the use of MT for assimilation (gisting) and MT for dissemination (post-edited MT). The growth in the amount of content to be translated and a push for cost-cutting from translation clients has meant that post-editing of MT has grown in popularity – a survey of almost 1000 language service providers (LSPs) in 2013 found that over 44% offer a post-editing (PE) service to customers (DePalma et al., 2013).

This has led to a requirement for user testing, as industry and researchers attempt to learn how translators work with MT, through the task of post-editing, and most usually within a translation memory tool (Moorkens and O'Brien, 2013). User dissatisfaction with post-editing has been widely reported (Krings, 2001; O'Brien and Moorkens, 2014) and translators tend to associate translation automation negatively with "regimentation, dependence, exploitation or impotence" (Cronin, 2013). Any new features intended to make the task more palatable to translators will naturally need to be tested for effectiveness. Automatic evaluation metrics (AEMs - such as BLEU) are typically used to measure quality improvements in MT and quality improvements, in turn, are expected to lead to higher levels of satisfaction among post-editors. However, some AEMs have been shown not to correlate well with human evaluation of quality (Tatsumi, 2009), and although automatic metrics measuring edit distance such as Human-mediated Translation Edit Rate (HTER) (Snover et al., 2006) have better correlations with human judgements (Snover et al., 2009), evaluations with real users are often necessary to gain a deeper understanding of the human/machine interaction and relationship. User evaluation also offers the possibility of eliciting valuable qualitative data, which can give insights into barriers for adoption and acceptance.

 This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of the CNGL (www.cngl.ie) at Dublin City University and by the FALCON Project (falcon-project.eu), funded by the European Commission through the Seventh Framework Programme (FP7) Grant Agreement No. 610879.

Many translation user studies are carried out using translation students, often out of necessity (Morado Vázquez et al., 2013) or convenience (Bowker, 2005). On the other hand, the common orthodoxy is that, where possible, it is best to evaluate using experts – professional translators – because they are more representative of the target user group for MT. In this paper we focus on the ramifications of using one user type over another for post-editing research. We do this by comparing the results of a post-editing user evaluation study using two sets of participants, one novice group (translation students) and one expert group (professional translators and post-editors). We have chosen translation students rather than lay or untrained volunteer translators (Mitchell, 2015) as our novice group, as students are more likely to be participants in research. The purpose of the user evaluation was to test smart post-editing features that had been programmed into a beta post-editing environment in order to test their effectiveness, although we do not report results from that test here. Instead, we focus explicitly on differences between the two user groups and on their suitability as research participants. Such differences are sometimes acknowledged but sidestepped when reporting research results.

The measurements collected during the evaluations were speed (measured in source text words per second), edit distance (measured using the Translation Edit Rate (TER) metric), attitudes to post-editing (collected via a survey), and user engagement (we measure the number of clicks on experimental features in the translation interface as a proxy for user engagement).

Yamada (2012) compared novice and professional translators and found productivity increases in both groups using post-editing, although the student group tended to make fewer edits. García (2010) found that his students preferred post-editing to human translation, which might make them a more favourable group for user testing.

Jääskeläinen (2010) notes that not all professional translators can be considered expert, as they may not produce good quality translations or may fall into an automatic routine when they work. Moreover, a translator may be an expert in a specific domain, and not at all expert in another. In addition, she suggests that experts may underperform for reasons such as "inflexibility, over-confidence, or bias" (Jääskeläinen, 2010). More generally, professional users have been found to exhibit resistance when faced with change due to a bias toward the status quo (Samuelson and Zeckhauser, 1988), or if they feel they have not been involved in the decision to change (Hirschheim and Newman, 1988). This outline of previous work suggests that the use of professional translators in post-editing research needs careful consideration because not all professional translators are equal.

2 Methodology

This research follows on from an earlier study that sought to identify PE-specific features that could be incorporated into editing environments to make the task more efficient for post-editors as described in Moorkens and O'Brien (2013). Five of those features were programmed into a beta PE environment (called "PEARL") and tested in this study (change gender, change number, change case, reject MT output, and copy source punctuation to target).

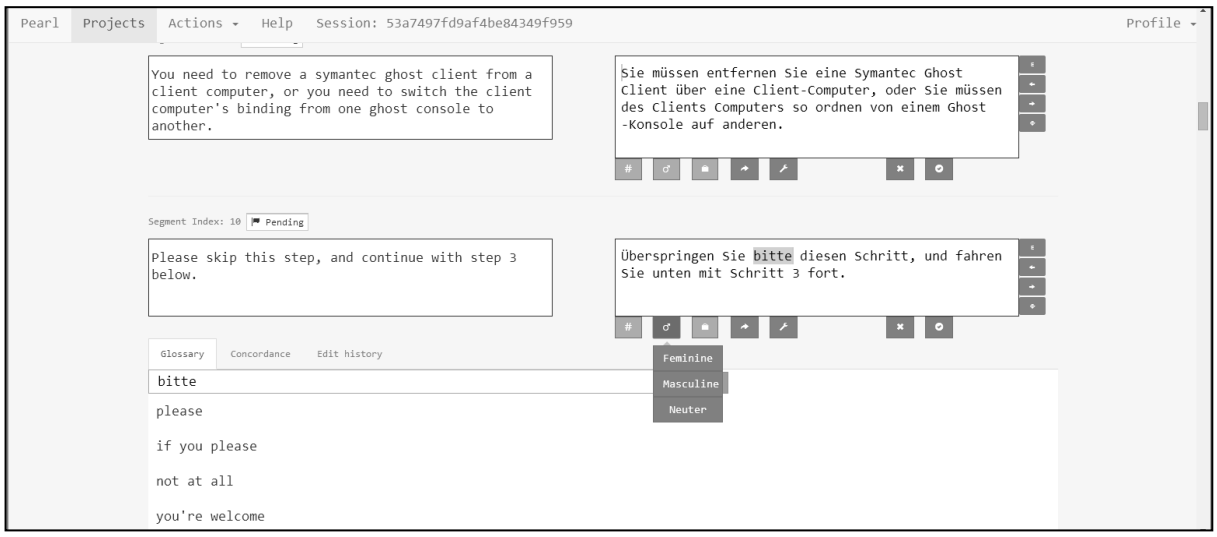

Figure 1. The PEARL test interface.

These features were selected because they represent some of the high-frequency, but tedious edits required during post-editing. They were tested in the English to German language pair using the purpose-built test interface with one group of professional and one group of student translators. English-German was selected because it is known to be one of the more demanding pairs for MT and we assumed we would see more evidence of issues regarding the features by using a demanding language pair.

2.1 Test Interface and Data

This research used the web-based interface called PEARL as a test suite for PE-specific functionality (see Figure 1). Data used were two test sets (50 US English segments each) of Norton Security helpdesk data, donated by Symantec, that had been machine translated into German using a purpose-built Moses Statistical MT engine. Features were switched on and off so that the two data sets could be tested with and without the new features.

Data Set 1 was post-edited by half of the participants with features turned off, and by half of the participants with features turned on. Then Data Set 2 was post-edited by half of the participants with features turned off, and by half of the participants with features turned on. Participants were requested not to switch applications, leave their desk, nor to ask any questions unless absolutely necessary. They were told not to worry about style, but to correct any words or grammar that was wrong or nonsensical. The researchers were present at all times during the post-editing sessions.

2.2 Participant Profiles

This research was carried out with two groups of participants. Group 1 was made up of nine expert participants, all professional English to German translators, mostly with extensive experience of localisation work who can intuitively translate and edit a text according to industry throughput expectations. In describing a five-stage process of gaining expertise, Dreyfus and Dreyfus (2005) highlight the importance of intuition as a defining characteristic of expertise. On average, the participants had 11.3 years of translation experience and four years of PE experience. Four participants had ten or more years' translation experience. The translators in Group 1 would regularly translate or post-edit texts similar to the data in this study, putting them at a further advantage when compared with the novice group, who had no experience of the specialised domain. The post-editing sessions took place in their normal place of work, on their usual computers.

Group 2 were 35 undergraduate translation students who were registered in an undergraduate translation programme in Zurich. The post-editing sessions took place in their computer lab. Very few had any professional translation experience, and the group were very reliant on procedural instruction, and as such could be considered novice according to Dreyfus and Dreyfus' taxonomy of expertise (2005).

Both groups of participants completed an online survey following the PE tasks, and the expert group also carried out a post-test interview. It was not possible to do so with the novice group due to timetable constraints.

2.3 Measurements

Participants were asked to undertake two post-editing tasks in English to German (one with the features to be tested and one without). The task comprised of 40 segments in total, although few of the novice participants completed the task within the allotted time (roughly 30 minutes per participant). From this task and from the post-task survey, we can compare our cohorts using four measurements. The first measurement is productivity or speed, which is calculated by dividing the number of words in the completed source text segments by total time in seconds, giving a words-per-second rate. The second measurement is edit distance, where raw MT and PE data are submitted to ASIYA[1], an online toolkit for MT evaluation (Giménez and Màrquez, 2010), to get a measurement using the Translation Edit Rate (TER) metric.

The third measurement is attitudes to post-editing. This was an open survey question that we have coded to a three-point Likert scale, where 1 is negative, 2 neutral, and 3 positive. More details of this coding phase are in Section 3.3. The final measurement is user engagement, looking at the number of times the participant clicked on experimental features in

[1] http://asiya.cs.upc.edu/demo/

the translation interface and using this as a proxy for user engagement. Participants were aware that feature-testing was the reason for the study, and were asked specifically to try the experimental features. Despite this, several participants chose not to try the features and post-edited as they would normally.

3 Results

3.1 Productivity

Table 1 shows the rate of source text words per second translated by Group 1, the professional post-editors, in two tasks (with/without new features). The average rate across all Group 1 users and tasks was 0.387 words per second after removing one outlier – User 2 was called away from his desk during the study, which made his second task time inaccurate and gave him a low WPS rate for that task (italicised). Table 2 shows the equivalent productivity rates for Group 2, the novice post-editors. The study with Group 2 was conducted in three university-scheduled computer lab sessions. For space reasons, we present the results for the first session of Group 2, with the average WPS rate (based on source text words translated) of 0.126. The figures for the rest of the group were very similar, with an average WPS rate across the whole group of 0.156, less than half the speed of the expert group. This is to be expected, of course, as the expert group have a great deal of experience in translation and in post-editing generally, as well as domain-specific expertise.

User	WPS Task 1	WPS Task 2
User 1	0.355	0.418
User 2	0.32	*0.109*
User 3	0.322	0.368
User 4	0.415	0.676
User 5	0.336	0.271
User 6	0.334	0.306
User 7	0.514	0.493
User 8	0.479	0.292
User 9	0.324	0.361
Average words per second for all users in both tasks		**0.387**

Table 1. Group 1 – Experts: Productivity (Words per Second)

User	WPS Task 1	WPS Task 2
User 1	0.072	0.117
User 2	0.136	0.118
User 3	0.129	0.103
User 4	0.148	0.157
User 5	0.210	0.129
User 6	0.151	0.115
User 7	0.091	0.151
User 8	0.087	0.129
User 9	0.127	0.106
User 10	0.240	0.130
User 11	0.052	0.091
User 12	0.057	0.080
User 13	0.202	0.137
Average words per second for these users in both tasks		**0.126**

Table 2. Group 2: Novices - Productivity (Words per Second)

3.2 Edit Distance

Using raw machine translated output and post-edited data, edit distance was calculated using the TER metric, defined by Snover et al. (2006, p3) as "the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references".

At the document level, the average TER score for Group 1 is 30.31, calculated by dividing the number of edits by the average number of words in the reference segment (the raw MT). The most heavily edited segment received a score of 122.22. The MT output and post-edited version of this segment may be seen in Table 3.

MT output	Post-edited segment
Bitte beachten Sie die Bedingungen in Ihrem Symantec-Supportzertifikat	*Informationen zu den Bestimmungen und Bedingungen der Vereinbarung finden Sie im Symantec-Support-Zert*

Table 3. Group 1 post-edit example

The novice post-editors in Group 2 tended to edit less, with an average document-level

TER score of 27.15. The most heavily edited segment, with a score of 100.0, may be seen in Table 4.

MT output	Post-edited segment
Microsoft hat einige Sicherheitslück beheben April einen Patch veröffentlicht.	*Microsoft hat im April einen Patch veröffentlicht, mit dem mehrere Sicherheitslücken behoben wurden.*

Table 4. Group 2 post-edit example

In making fewer edits, Group 2 left more errors in the raw MT uncorrected. For example, in the segment "*Es tut mit leid, aber Ich kann bei diesem Produkt nicht weiter assistieren*", the post-editor has left the misspelled *tut mit leid* unedited, whereas all of Group 1 corrected this phrase to *tut mir leid*. Group 2 target texts contained more misspellings, such as the word *kann* spelled with a single 'n'.

3.3 Attitude to post-editing

Responses to the question 'Did you like the task of post-editing? Why/why not?' were divided into positive, neutral and negative. A response was categorised as positive if the participant answered with responses such as "Yes", "I liked it", or "it was kind of fun", neutral if they used phrases such as "so, so", "sort of", "kind of" or if they used some form of neutral description, and negative if they said "no", "not really", or "I think it is a bit useless". Comparative responses by group may be seen in Table 5.

	Group 1 (experts)	Group 2 (novices)
Positive	11%	35%
Neutral	33%	18%
Negative	56%	47%

Table 5. Attitudes to post-editing

When asked for their views on post-editing prior to the evaluation, Group 1 responses were mostly negative. Three participants said that PE can be worthwhile if the MT quality is good enough. Others disliked PE for reasons such as the lack of creativity, tediousness of the task, limited opportunity to create quality, poor quality source text rendering MT unusable, and poor term management. They considered that the main tasks during PE are tedious fixes to the word order, correcting product names, and correcting tags. They also said that they are more prone to mistakes as their "mind falls asleep", that they quickly become tired due to having to be constantly vigilant and due to the absence of any confidence indication, and that switching between mouse and keyboard was also tedious. They sometimes find it difficult to understand how to balance time and quality to find an acceptable quality level for a client.

In comparison, the novices in Group 2 were more positively disposed towards post-editing. Of those who gave positive responses, the reasons they used were that the translation was already done for them and they just needed to "improve a few things". Others liked the task because it was "new" or "challenging". Those with a neutral attitude suggested that post-editing limited the use of "imagination" or that it was "uncreative". Reasons given for negative responses can be grouped into four main categories to do with time, quality, tool functionality, and lack of context. Some participants complained about the raw MT quality saying it would be "easier to start from scratch". There was a perception among a few that the task took more time (than translation), was exhausting because it was repetitive, and made more difficult due to the lack of context for the segments.

3.4 User engagement

Participants were expressly requested to try several experimental features in the PEARL interface, but not all participants chose to engage with them. All were told that, as per DCU research ethics guidelines, they would not be penalised for non-participation, but all chose to participate. It is possible that they felt compelled by management or co-workers (in the case of Group 1) or lecturers and fellow students (in the case of Group 2). By taking part without engaging with the purpose of the research, a participant's impact is more negative and wasteful than not taking part at all. The average number of button presses on experimental features are shown in Table 6.

	Group 1 (experts)	Group 2 (novices)
Change case	2.50	7.26
Change gender	2.66	3.07
Change number	1.66	2.89

Table 6. Engagement with PE features

As can be seen, the experts in Group 1 were less likely to engage with the interface. The average number of button presses was brought down by two participants who chose not to try any of the buttons at all. All participants from Group 2 tried the feature buttons at least once, and most continued to engage with the purpose of the research despite some server problems causing an intermittent response to buttons pressed. As previously stated, one characteristic of an expert is intuition. Group 1 participants intuitively knew how to work quickly on an MT segment using familiar features (such as cut and paste), but this made them less likely to try unfamiliar features, such as those added for the purpose of this research.

4 Conclusion

User evaluation is currently continuing on post-editing with foci on areas such as adding PE-specific features (Sanchis-Trilles et al., 2014), incremental retraining (Dara et al., 2014), deciding what content should be post-edited rather than translated from scratch (Castilho et al., 2014), quality prediction (Vieira, 2014), and quality/productivity expectations in an MT/TM combination (Guerberof Arenas, 2014). Results of these evaluations may have an impact on decisions as to what remuneration is appropriate for professional post-editing. As MT deployment increases in the language industry, it makes sense to carry out user evaluations with the people who will be expected to engage with that technology. Productivity rates for experts, as seen in Section 3.1, were more than double those of the novice post-editors. In fact, the expert post-editors in Group 1 of this study worked so quickly that our server's CPU load rose worrisomely as they moved quickly and intuitively through the texts. Their segments tended to be more comprehensively edited than those of the novice group. On the other hand, their attitudes towards the technology were considerably more negative than that of the novice group and they were much more likely to adhere to an automatic routine, and less likely to engage with the research objectives. Their attitudes are possibly due to "anxiety and uncertainty regarding change" (Kim and Kankanhalli, 2009).

It is unclear whether the lower engagement with the research (in Section 3.4) by the expert group was due to their automatic routine or a negative attitude to PE/MT, but it appears that novice users are more likely to engage with new tasks and features without preconceptions. It must also be noted that, despite the comparatively positive attitude to PE among the novice group, almost half still felt negatively about the task of PE. The novice group was enthusiastic about taking part in research, and as with research in general, student groups are likely to take part in future research due to convenience and lower costs. This research suggests that, for post-editing, there are tradeoffs to be considered when using novice vs. professional groups to estimate productivity or the usefulness of a new feature in a production environment. Novices may generally be more positive and enthusiastic and will engage considerably with the research, but conclusions drawn from research with novice users cannot necessarily be carried over to experts. Professionals will be more efficient, but their routines and attitudes may prevent full engagement with research objectives. To get balanced results on user interaction with MT, it is advisable to employ adequate numbers of users with varying levels of expertise.

Acknowledgement

The authors would like to place on record their thanks to the staff and management at Alpha CRC in Cambridge, UK, and at ZHAW in Winterthur, Switzerland, for their help and participation in this research. We also thank Chris Hokamp and Ximo Planells for development work, and Dr. Lamia Tounsi and Peter Jud for assistance in Winterthur. In addition, we are grateful to Symantec for providing test data.

References

Bowker, Lynne. 2005. Productivity vs quality? A pilot study on the impact of translation memory systems. *Localisation Focus*, 4(1):13-20.

Castilho, Sheila, Sharon O'Brien, Fabio Alves, Morgan O'Brien. 2014. Does post-editing increase usability? A study with Brazilian Portu-

guese as Target Language. In *Proceedings of the Seventeenth Annual Conference of the European Association for Machine Translation (EAMT2014)*. Proceedings eds. Marko Tadić, Philipp Koehn, Johann Roturier, Andy Way. Dubrovnik, Croatia, June 16-18 2014, 183-190.

Cronin, Michael. 2013. *Translation in the Digital Age*. Routledge, Oxfordshire, UK.

Dara, Aswarth, Josef van Genabith, Qun Liu, John Judge, Antonio Toral. 2014. Active Learning for Post-Editing Based Incrementally Retrained MT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, April 26-30 2014, 185–189.

DePalma, Donald A., Vijayalaxmi Hegde, Hélène Pielmeier, and Robert G. Stewart. 2013. *The Language Services Market: 2013*. Common Sense Advisory, Boston, USA.

Dreyfus, Hubert, and Stuart Dreyfus. 2005. Peripheral Vision: Expertise in Real World Contexts. *Organization Studies,* 26 (5): 779–792.

Garcia, Ignacio. 2010. Is machine translation ready yet? *Target,* 22(1):7-21.

Giménez, Jesús, and Lluís Màrquez. 2010. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, 94:77-86.

Guerberof Arenas, Ana. 2014. Correlations between productivity and quality when post-editing in a professional context. *Machine Translation*, 28(3-4): 165-186.

Hirschheim, R., and M. Newman. 1988. Information systems and user resistance: theory and practice. *The Computer Journal*, 31(5):398-408.

Jääskeläinen, Riitta. 2010. Are All Professionals Experts? Definitions of Expertise and Reinterpretation of Research Evidence in Process Studies. In *Translation and Cognition*, ed. by Gregory Shreve and Erik Angelone. John Benjamins, Amsterdam, Netherlands 213–227.

Kim, Hee-Woong, and Atreyi Kankanhalli. 2009. Investigating User Resistance To Information Systems Implementation: A Status Quo Bias Perspective. *MIS Quarterly*, 33(3):567-582.

Krings, Hans P. 2001. *Repairing Texts*. Kent State University Press, Ohio, USA.

Mitchell, Linda. 2015. The potential and limits of lay post-editing in an online community. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT2015)*, Antalya, Turkey, May 11-13 2015.

Moorkens, Joss, and Sharon O'Brien. 2013. User Attitudes to the Post-Editing Interface, In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*. Proceedings eds. Sharon O'Brien, Michel Simard, Lucia Specia. Nice, September 2, 2013, 19–25.

Morado Vázquez, Lucía, Silvia Rodríguez Vázquez, and Pierrette Bouillon. 2013. Comparing Forum Data Post-Editing Performance Using Translation Memory And Machine Translation Output: A Pilot Study. In *Proceedings of MT Summit XIV*. Nice, France, September 3-6, 2013.

O'Brien, Sharon, and Joss Moorkens. 2014. Towards Intelligent Post-Editing Interfaces. In *Proceedings of FIT XXth World Congress 2014*, 4-6 Aug 2014, Berlin, Germany.

Samuelson, William, and Richard Zeckhauser. 1988. Status Quo Bias in Decision Making. *Journal of Risk and Uncertainty*, 1:7-59.

Sanchis-Trilles, Germán, Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Jesús González-Rubio, Robin L. Hill, Philipp Koehn, Luis A. Leiva, Bartolomé Mesa-Lao, Daniel Ortiz-Martínez, Herve Saint-Amand, Chara Tsoukala. 2014. Interactive Translation Prediction vs. Conventional Post-editing in Practice: A Study with the CasMaCat Workbench. *Machine Translation*, 28(3-4):217-235.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*, 2006, 223–231.

Snover, Matthew, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. 2009. Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, 30-31 March 2009, 259–268.

Tatsumi, Midori. 2009. Correlation between automatic evaluation scores, post-editing speed and some other factors. In *Proceedings of MT Summit XII*, Ottawa, 26–30 August 2009, 332–339.

Vieira, Lucas Nunes. 2014. Indices of cognitive effort in machine translation post-editing. *Machine Translation*, 28(3-4):187–216.

Yamada, Masaru. 2012. Revising text: An empirical investigation of revision and the effects of integrating a TM and MT system into the translation process. *PhD Thesis*, Rikkyo University, Tokyo.