This translation is not too bad: An analysis of post-editor choices in a machine translation post-editing task

Maarit Koponen
University of Helsinki, Dept of Modern Languages
P.O. Box 24
00014 Helsingin yliopisto
maarit.koponen@helsinki.fi

Abstract

Variation between post-editors of machine translation is a well-known issue. This variation shows itself in post-editing speed, amount of editing and differing final translations. However, relatively few studies exploring the differences have been reported. This paper describes a post-editing task involving controlled language tourist phrases translated from English into Finnish. Post-editors select the best out of three machine translated suggestions, which they can accept without editing or post-edit as necessary. Agreement between editors is analyzed and reported in terms of selecting the best suggestion, deciding its acceptability, and producing a final post-edited version. Editors are compared in terms of post-editing time, edit distance and final translations created. With a qualitative analysis, we examine differences between the selected and rejected suggestions as well as differences between the post-edited versions created by different editors. Examples of editor preferences are also discussed.

1 Introduction

The growing interest in, and use of, machine translation (MT) post-editing as a way to increase productivity in professional translation scenarios has also recently led to growing interest on the research side. Tools and practices for post-editing (PE) are being developed, and PE tasks are being used to evaluate MT quality. A recognized issue in post-editing scenarios is the variation between different editors, which shows in the amount of editing, PE speed and differing final translations produced. Post-editing, like translation in general, is an inherently subjective task in that the source meaning can generally be expressed in the target language in more than one way.

In analysing the variation between post-editors, attention has generally focused on questions of productivity: PE time and the technical effort of post-editing measured as keystrokes or edit distance between the MT and PE version (Krings, 2001; Plitt and Masselot, 2010; Tatsumi and Roturier, 2010; Koponen et al., 2012). Some studies have included analysis of the numbers of PE versions created and PE versions preferred by evaluators (Tatsumi et al., 2012), or examples of differing PE versions (O'Brien, 2005). However, much of the variation in post-editor choices and preferences as well as the factors influencing these choices still remains to be investigated.

In this paper, we aim to take some steps toward exploring the variation between editors in terms of the amount of editing performed (number of sentences edited and edit distance) and PE speed. We also examine the agreement between editors in selecting the best MT suggestion and deciding on whether to edit or not. For this purpose, we analyze data collected during a post-editing task involving a multilingual, controlled language generation and machine translation tool. The material, generated according to the controlled language rules, consists of short, relatively simple ”tourist phrases” with limited vocabulary and structures: for example, questions about prices and directions, or small talk phrases. This type of material was selected for this study for its simplicity. The short, controlled sen-

Sharon O'Brien, Michel Simard and Lucia Specia (eds.)
Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice, Nice, September 2, 2013, p. 1–9.

tences were expected to lead to a relatively small number of MT errors, which would decrease the need for extensive rewriting and help to isolate the post-editing choices by different editors.

The post-editing task described in this paper involves the editors selecting the best out of three MT suggestions, accepting it without modification or post-editing as necessary. With this data, we set out to investigate the choices made by the editors in which suggestion to select and whether to accept it as such or edit. We will examine agreement between the editors and variation between different editors as well as the different PE versions. Using qualitative analysis, we will examine some of the preferences shown by the editors.

This paper is organized as follows. Section 2 presents prior research related variation in MT suggestion selections and post-editing. Section 3 describes the material and methods used in the analysis. Section 4 presents the analysis results. Conclusions from this study as well as future work are discussed in Section 5.

2 Related work

Selection of the best MT suggestion has often been used in large MT evaluation campaigns, such as those organized in context of the annual Workshop on Statistical Machine Translation (WMT) (Callison-Burch et al., 2012), where evaluation of competing systems involved the ranking of alternate MT suggestions. The results mostly focus on evaluating the quality of different systems, and do not generally include analysis of the differences between high and low ranking suggestions.

Prior MT post-editing studies have included analyses of the differences in PE speed and edit distances between different post-editors. An extensive analysis of increased productivity in post-editing compared to translation was carried out by Plitt and Masselot (2010) in a study involving twelve professional translators and various language pairs. Significant variation in post-editing speed was observed between post-editors.

In a study involving nine professional translators post-editing English-to-Japanese MT, Tatsumi and Roturier (2010) found that the translators differed more in terms of editing time than textual changes.

Koponen et al. (2012) report an analysis of human variability in post-editing. Eight post-editors editing English-to-Spanish MT were compared in terms of PE time, keystrokes during editing, and edit distance. Post-editors were found to differ more in terms of PE time and keystrokes than edit distances.

Tatsumi et al. (2012) report the frequencies of multiple successive PE versions occurring in a scenario where post-edited versions of MT in various language pairs are crowdsourced from student participants. Most sentences are found to have no more than one translation version. Different crowdsourced versions are subsequently evaluated by professional translators who could either accept or revise them, and results are reported comparing whether the last or some earlier PE version is accepted.

One approach to studying different translation or post-editing choices is Choice Network Analysis (CNA). CNA has been suggested by Campbell (2000a,b) as a way to compare different translation versions created by multiple translators for a given source string. In Campbell (2000b), CNA is used to examine the behaviour of nine students related to two specific structures (cross-clause ellipsis and relative clauses).

O'Brien (2005) presents examples of using CNA to analyze the translations of four translator students post-editing English-to-German MT. Results obtained with CNA are compared to post-editing data, and long pauses in editing are found to correlate with locations indicated as difficult by CNA.

Blain et al. (2011) introduce the concept of Post-Edit Action (PEA), which combines multiple edit operations to linguistically logical groups, and analyze post-editing changes made by four professional translators on English-to-French MT. No comparison of individual editors' choices are reported.

The purpose of this study is to explore some aspects of agreement or disagreement between post-editors. Rather than post-editing times and edit distances, we focus on the agreement in selecting the best translation suggestion and in deciding acceptability. Further, we examine differences between individual editors and the final PE versions they create.

3 Material and analysis methods

The material analyzed for this paper consists of 139 sample sentences and their translations obtained from the pilot evaluation material of a multilingual, controlled language text generation and machine translation system developed as part of the European MOLTO project[1]. The dataset, described in more detail in Rautio and Koponen (2013), includes English source sentences, three Finnish MT versions of each sentence, and post-editing data from 11 post-editors. The total number of source words is 827, and source sentence length varies from 2 to 15 words per sentence (median 5).

The MT versions have been produced with the rule-based generation and translation tool in question, as well as the statistical MT systems Google Translate[2] and Bing Translator[3]. In some cases, multiple systems had produced the same MT suggestion. All systems produced an identical suggestion for 8 sentences, and two systems produced identical suggestions for 41 sentences. For 90 out 139 sample sentences, three different MT suggestions were provided.

The post-editing data was collected using the open-source online MT evaluation tool Appraise (Federmann, 2012). The sentences were post-edited by 11 translator students who were native speakers of Finnish. Each editor was shown the 139 English source sentences together with the three Finnish MT suggestions. The order of sentences and suggestions was randomized by the evaluation tool. The editors were asked to select the MT suggestion they considered best and accept it as-is or post-edit as necessary. They were instructed to make only minimal corrections necessary. The option to create a translation from scratch was also given. In total, the dataset contains 1527 final translations created by the editors either by accepting or editing the MT suggestions. No sentences had been translated from scratch.

As the original sample sentences had been generated for the testing the rule-based system that was used to produce one of the MT versions, suggestions by this system can be expected to have an advantage in the selection. However, for the purposes of this study, we are interested in cases of agreement or disagreement between editors rather than the relative success of the systems.

The analysis of agreement between the editors involves two issues: whether they agree on the selection of the best MT suggestion, and whether they accept the suggestion as-is or edit it. Combining these aspects leads to the following six possible scenarios:

1. The same MT suggestion is selected by all.
   All accept without editing.

2. The same MT suggestion is selected by all.
   None accept without editing.

3. The same MT suggestion is selected by all.
   Some accept without editing.

4. Different MT suggestions are selected.
   All accept without editing.

5. Different MT suggestions are selected.
   None accept without editing.

6. Different MT suggestions are selected.
   Some accept without editing.

Using the collected post-editing data, all sentences were categorized according to these scenarios. The final PE versions created by the editors were compared to calculate the number of different versions created for each source sentence. The most common PE version was also recorded.

Differences between individual editors were examined in terms of the amount of editing, PE time and agreement with the most common choices across all editors. For comparing the amount of editing, the Human-targeted Translation Edit Rate (HTER) was calculated using TERplus (Snover et al., 2009). The HTER score is calculated as the number of edit operations (word insertions, deletions, substitutions or word order shifts) between the MT and PE version divided by the number of words in the PE version. A HTER score of 0 indicates no changes while 1 indicates complete rewriting.

Sentence-level PE time automatically recorded by the evaluation tool was used for time comparisons. Information about the editors' choice of MT suggestion was compared to the most common selection for each sentence. Similarly, the editors' final version was compared to the most common version for each sentence.

Finally, the MT suggestions and final versions were analyzed manually by a native Finnish speaker. To investigate why some MT suggestions

[1] http://www.molto-project.eu/

[2] http://translate.google.com

[3] http://www.bing.com/translator

were preferred over others, the selected and rejected MT suggestions were assessed for the correctness of meaning and language on a strict binary scale (fully correct/not fully correct). Cases where multiple PE versions had been created were compared to examine the differences between these versions.

4 Analysis results

This section presents the analysis results. Overall agreement between editors in terms of the MT suggestion selected and choice to accept or edit is presented in Section 4.1. The comparison of individual editors is presented in Section 4.2. The qualitative analysis of differences in MT suggestions and PE versions are discussed in Section 4.3.

4.1 Agreement between editors

	MT suggestions selected		
	Same	Different	Total
All accept	44	15	59
None accept	1	5	6
Some accept	33	41	74
Total	78	61	139

Table 1: Number of sentences categorized according to editor selections of same or different MT suggestions and choice to accept or edit.

Table 1 shows the distribution of cases into the six selection/acceptance categories. The columns show the number of sentences categorized by whether all editors selected the same MT suggestion or whether different suggestions were selected, as well as the total. The rows show the number of sentences categorized by whether the suggestion chosen by each editor was accepted by all, none or some editors.

Overall, the editors appear to mostly agree on which suggestion they select. When the 8 cases with three identical MT suggestions are excluded, all 11 editors select the same MT suggestion for 70 out of the remaining 131 source sentences (53%). In a further 29 cases (22%), only one editor selects a different option. This leaves 32 sentences (24%) where two or more people select a different suggestion. Only one case was found where each of the three MT suggestions were selected as best by at least one editor.

When the editors agree on the same MT suggestion, it is most often (44 sentences, 56.4%) accepted without editing. The cases where all editors found some suggestion acceptable, but disagreed on which one, were less common. For these 15 sentences, it appears that two of the suggestions are correct although different. Overall, there were only 6 cases where none of the editors found any MT suggestion acceptable. In one case they agree on which is the easiest to correct, whereas in the other five cases, different MT suggestions are selected.

The remaining cases represent a mixed situation where some accept and some edit. When one MT suggestion is selected by all (33 sentences), this suggestion still appears to be superior, but the editors disagree on whether it can be accepted as such. On the other hand, when the editors select different MT suggestions with some accepting and others editing (41 sentences, 67.2% of the cases where selections of best suggestion are split), there seems to be even more disagreement: some are willing to accept one suggestion while others rather edit a different one. Some potential reasons for these preferences are discussed in Section 4.3.

Table 2 shows the numbers of different PE versions produced for a given source sentence (1, 2, 3 or more than 3 versions). In addition to the total number of sentences, the rows show the number of sentences divided into the six defined selection/acceptance scenarios.

		Number of PE versions				
		1	2	3	≥ 4	Total
Same MT	All accept	44	0	0	0	44
	None accept	0	0	0	1	1
	Some accept	0	21	7	5	33
Different MT	All accept	0	15	0	0	15
	None accept	1	0	3	1	5
	Some accept	3	11	13	14	41
Total		48	47	23	21	139

Table 2: Number of different PE versions created by editors in the six selection/agreement scenarios.

Overall, most of the 139 sentences have only one or two final PE versions. For 48 sentences, only one final version was found. Nearly all of them (44 sentences) naturally relate to the cases where all editors have accepted the same version without modification. In one instance, all editors ended up with the same PE version although one of them started with a different MT suggestion, and in three cases, one editor chose to edit a different suggestion but produced a PE version identical to the MT suggestion that was accepted by the other editors.

Cases with two different final versions were mostly produced by some editors accepting and others editing the same MT suggestion (21 sentences) or different editors accepting different suggestions (15 sentences). The remaining 11 cases involve situations where different suggestions are selected with varying acceptance or editing. For sentences with more than two versions, most also result from different suggestions being selected and varying choices whether to accept or edit. The highest number of different PE versions found was 10 (1 sentence).

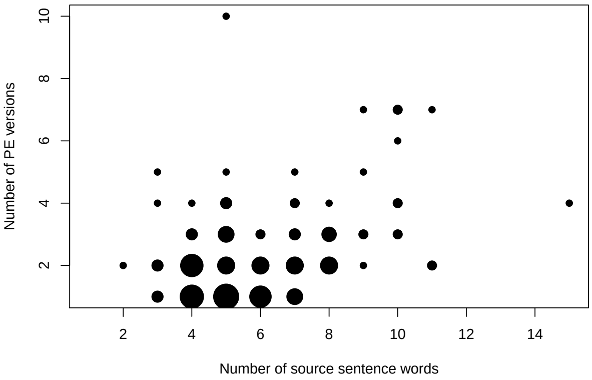

Figure 1: Plot showing the number of PE versions plotted against the number of source words. The circle size indicates multiple sentences with same values.

Figure 1 shows the number of PE versions for each sentence plotted against the number of source sentence words, with larger circles representing multiple sentences with the same value. All 48 cases with only one PE version, and nearly all with two versions (37 out of 47), involve sentences with 7 words or less. Most longer sentences, on the other hand, have 3 or more PE versions. With the exception of the one 5-word sentence with 10 different versions, sentences with the highest number of PE versions have more than 8 words. This is likely at least partly due to the shorter sentences having better MT quality and more often being accepted as-is. Conversely, longer sentences contain more errors and more need for editing then leads to more variation in the solutions found by different editors.

4.2 Comparison of individual editors

Figure 2 shows the number of sentences edited by each editor. Overall, all editors mostly accept one of the suggestions as-is. The number of sentences edited by each editor ranges from 19 (13.7% of all sentences, FI11) to 46 (33.1%, FI01) with a median of 39 sentences (28.1%).

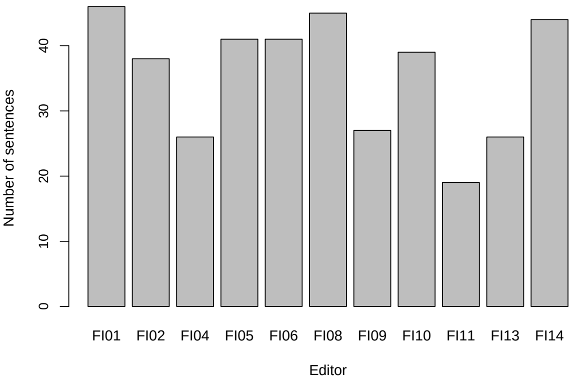

Figure 2: Bar plot showing the number of sentences edited by each editor.

Figures 3(a) and (b) show box plots of edit distances by editor. In the box plots, the bottom and top of the box represent the first and third quartile, and the line inside the box shows the median. The whiskers extend to 1.5 times the length of the box, and individual circles represent the cases outside of these limits.

Figure 3(a) shows the edit distance for all sentences. Because all editors accepted the majority of sentences without editing (see Figure 2), the median HTER score for each editor 0 is in Figure 3(a). To provide a clearer picture of how much each editor edited when they *did* decide editing was necessary, Figure 3(b) shows the edit distances for only those sentences that had been edited. The low HTER scores indicate that even when editing is considered necessary, a relatively small number of changes is made. Overall, there do not appear to be great differences between the editors, as all editors have median HTER between 0.17 and 0.20 except FI04 (median 0.23).

Figure 4 shows box plots of PE times by editor. One outlier sentence (from editor FI13) with a PE time over 300 seconds was removed, which was the only case where PE time exceeded 100 seconds. Median times are between 5.4 and 10.0 seconds per sentence for all editors except

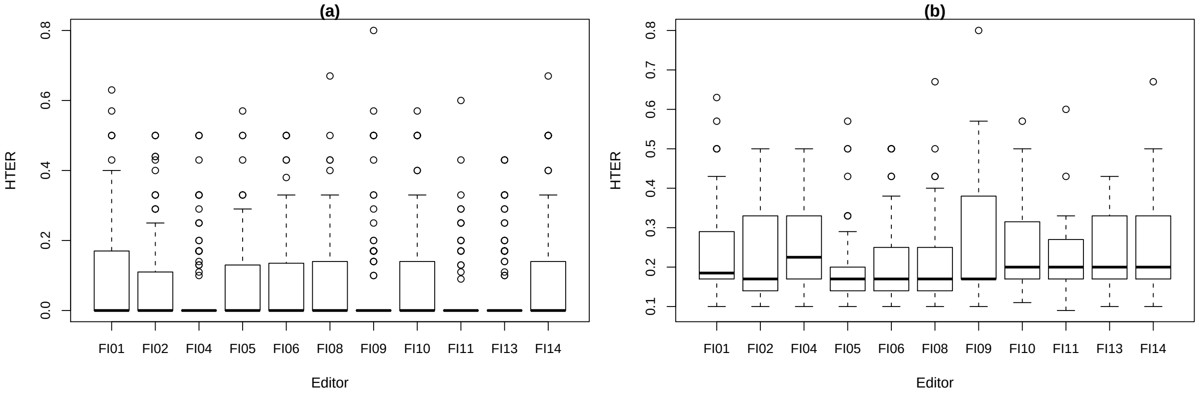

Figure 3: Box plots showing the edit distances (HTER) for each editor. HTER scores are shown for all sentences (a) and for edited sentences only (b).

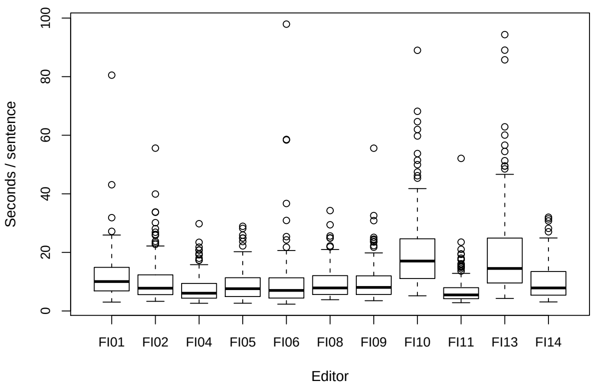

Figure 4: Box plots showing the editing times for each editor.

FI13 (14.5 seconds/sentence) and FI10 (17.0 seconds/sentence). For the three fastest editors (FI04, FI05, FI11), even the slowest times are around 30 seconds per sentence.

The editors were also compared in terms of how often their choice of best MT suggestion differed from the majority and how often they produced a final PE version differing from the most common version.

Figure 5 shows a bar plot of the number of sentences where each editor selected a different MT suggestion than the majority. The number of cases where each disagreed with the majority varies between 6 (FI01) and 17 (FI13), with median of 10.

Figure 6 shows a bar plot of the number of cases where each editor produced a PE version different from the majority. The number of such differing versions ranges from 19 (FI09) to 35 (FI02), with a median of 29.

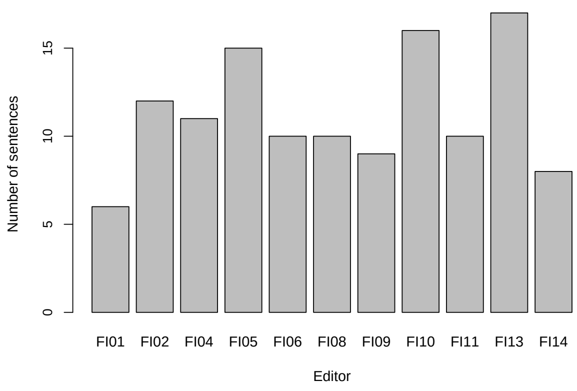

Figure 5: Bar plot showing the number of sentences where each editors' selection of MT suggestion differs from the majority.

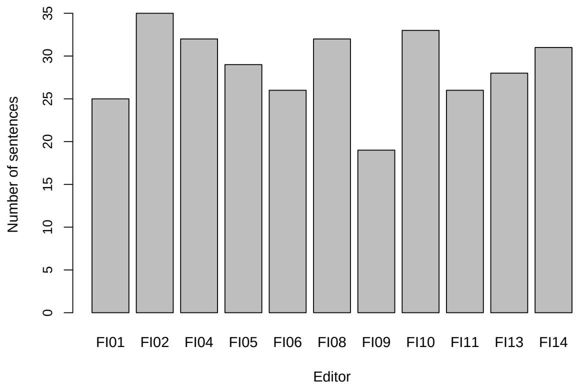

Figure 6: Bar plot showing the number of sentences where each editors' PE version differs from the majority.

Comparing these figures, some editors appear to stand out: Editor FI04 has one of the smallest numbers of edited sentences and is also one of the fastest, but seems to edit slightly more than the others and is among those who most commonly produce a PE version differing from the majority. On the other hand, editor FI11 edits relatively few sentences, but is one of the slowest and rarely deviates from the most common PE version. Editors FI01 and FI09 appear to commonly agree with the majority, as both have low numbers of selections and PE versions differing from the majority. These observations of differing profiles share some similarities with results reported in Koponen et al. (2012).

4.3 Qualitative analysis of editor preferences

Section 4.1 presented results of how often the editors agreed in selecting the best MT suggestion and whether it needed further post-editing. In this Section, we aim to examine some possible explanations for their choices and differences by comparing, on one hand, the selected and rejected MT suggestions, and on the other hand, the differing PE versions.

In the cases where at least two MT suggestions were offered and all editors selected the same one (70 sentences), most rejected suggestions had neither correct meaning nor correct language. They generally contained multiple errors which sometimes made it difficult to ascribe any meaning to the sentence – for example, *Jossa on suosituin Puolan ravintolassa?* 'In which is the most popular of Poland in the restaurant?' for *Where is the most popular Polish restaurant?* Interrogative sentences commonly contained multiple MT errors involving missing interrogative suffixes and wrong word order, literal translations of *do* or incorrectly added negation, affecting both language an meaning.

There were 15 cases where the rejected MT suggestion had correct language but different meaning than the source. This occurred, for example, in possessive structures, where the wrong case in the possessor noun can lead to suggestions like *Hänen vaimonsa on maitoa.* 'His wife is milk' instead of *Hänen vaimollansa on maitoa.* 'His wife has milk.' Other changed meanings involved incorrect words. In 3 cases, the rejected MT suggestion was assessed to have correct meaning despite incorrect language. These involved sentences with incorrect subject-verb agreement that is not standard in written Finnish but commonly used in spoken language.

In 3 cases, the meaning and language of the rejected suggestions was correct, but all the editors still preferred another suggestion. In these cases, the editors appear to have made the decision based on specific words or expressions, such as *Oletko kahdeksanvuotias?* for 'Are you eight years old?' rather than *Oletko kahdeksan vuotta?* Other sentences with similar expressions and varying editor choices were found.

In the 15 cases where all editors have accepted some MT suggestion without editing but disagree on which one, the selected suggestions generally differ from each other in ways that leave both the meaning and language correct. They mostly involved choice between synonymous words or expressions, such as *avoinna* vs *auki* for 'open'. As Finnish word order is relatively free, word order was also a recurring difference. One case of differing punctuation was also found.

In 4 cases, one of the selected versions had correct language but was not, in fact, precisely correct in terms of meaning. These sentences involved cases where Finnish makes a distinction not present in English: the pronoun *they*, where Finnish uses different words for humans and non-humans, or second person forms, where Finnish distinguishes between informal singular, polite singular, and plural. During the post-editing task, the English sentences were presented with disambiguation information, but some editors appear to have ignored this. Two cases where MT suggestions with both incorrect meaning and incorrect language were accepted by at least one editor were also found.

Differences in preferences could also be observed in the cases where the editors disagreed whether the same MT suggestion needed editing or not and produced differing PE versions. Some recurring differences involved punctuation, specifically commas between main and subordinate clauses (required, but commonly omitted particularly in short sentences), alternate spellings such as *pizza* vs *pitsa* 'pizza' or alternate suffixes such as *dollareja* vs *dollareita* 'dollar (plural partitive)', as well as synonyms.

At least some cases where the editors disagree on which MT suggestion to select and whether it needs editing appear to be connected to particularly strong preferences for specific words or expressions. One such preference involved the choice of the verb *tahtoa* or *haluta* 'want'. Most editors appear to have at least some preference for *haluta*, since options containing that word were generally selected by all or nearly all editors if otherwise correct, and MT suggestions containing the alternative *tahtoa* were often edited to change this verb even when otherwise correct. Seven cases were identified where at least one editor even chose to edit MT suggestions that had both incorrect language and different or unclear meaning but contained a form of the preferred verb *haluta*, rather than accept or even edit a (correct) version containing *tahtoa*.

Some sentences or expressions seemed to generate a large number of different PE versions. As mentioned in Section 4.1, one sentence received a total of 10 different versions: *This apple is not too bad*. Without context, it can be understood either literally (”bad, but not too much so”) or idiomatically (”quite good”). Other sentences with the ”not too (adjective)” structure have been interpreted literally by the editors, but in this case, all but one chose the idiomatic interpretation and expressed it with varying wording. Other cases leading to multiple PE versions involved sentences like *Do you know how far the park is by bike?* for which a total of seven different versions were created.

Such cases where particularly many versions were created are similar to findings obtained using Choice Network Analysis. CNA assumes that multiple target versions of a given source string indicate parts that are difficult cognitively, as no single obvious solution is available to the translator or post-editor (Campbell, 2000b). A connection with pauses during post-editing was reported in O’Brien (2005), supporting this assumption. The sentences involving *not too bad* and *how far by* may indeed have caused difficulty. However, for some of the differences, such as the word choice for translating *open*, the variation may simply indicate varying preferences without any particular difficulty.

5 Conclusions and future work

The purpose of this study was to examine editors involved in an MT post-editing task, their editing choices and agreement between editors.

For most source sentences, all or all but one editor select the same MT suggestion and most sentences only have one or two PE versions. This is likely to be related to the nature of the controlled language, high MT quality, and the large number of suggestions accepted without modification. Some differences were found between individual editors in terms of the number of sentences edited, and how often they deviated from the most common selection or most common PE version. Similar to prior studies, less variation was observed in edit distances than in PE times.

As expected, the editors tended to reject MT suggestions with multiple errors leading to both incorrect meaning and language. Variation in the selection of best MT suggestion and final PE versions appeared to mainly relate to choice of specific words or expressions or the use of punctuation. Cases where some editors chose to edit an incorrect sentence over a version accepted as correct by others were also identified. Examples of editor preferences related to these choices were discussed.

The sample is rather limited due to the controlled language. However, the repeated vocabulary and structures offer a chance to observe editor choices across similar cases. In future work, we are interested performing a more quantitative analysis of factors potentially influencing the editors’ choices. Working with a less controlled text type would likely reveal more variation in the editor’s choices, and would therefore be desirable. Professional translators might also produce results different from translator students. One question to study would be whether the version containing the preferred words (even in incorrect form) or word forms (even if not preferred words) would be preferred by editors. For this purpose, automatic tagging and lexical resources such as WordNets could be used. Specific editor preferences could also be explored in more detail.

Acknowledgements

This work has been supported by LANGNET Finnish Graduate School in Language Studies. The author wishes to thank the EU MOLTO project, particularly Evaluation Coordinator Jussi Rautio and Professor Lauri Carlson, for the use of the evaluation material and for discussions.

References

Blain, Frédéric, Jean Senellart, Holger Schwenk, Mirko Plitt and Johann Roturier 2011. Qualitative analysis of post-editing for high quality machine translation. In *MT Summit XIII: the Thirteenth Machine Translation Summit*, Xiamen, China. 164–171.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut and Lucia Specia 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montréal, Canada. 10–51.

Campbell, Stuart. 2000. Choice Network Analysis in Translation Research. In M. Olohan (ed.) *Intercultural Faultlines: Research Models in Translations Studies: Textual and Cognitive Aspects*. St. Jerome, Manchester. 29–42

Campbell, Stuart 2000. Critical Structures in the Evaluation of Translations from Arabic into English as a Second Language. *The Translator*, 6: 37–58.

Federmann, Christian. 2012. Appraise: An Open-Source Toolkit for Manual Evaluation of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 98: 25–35

Koponen, Maarit, Wilker Aziz, Luciana Ramos and Lucia Specia. 2012. Post-editing time as a measure of cognitive effort. *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP 2012)*, San Diego, USA. 11–20.

Krings, Hans P. 2001. *Repairing texts: Empirical investigations of machine translation post-editing process*. G. S. Koby (ed.). The Kent State University Press, Kent, OH.

O'Brien, Sharon. 2005. Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability. *Machine Translation*, 19(1):37–58.

Plitt, Mirko and Francois Masselot. 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, 93: 7–16.

Rautio, Jussi and Maarit Koponen. 2013. *D9.2 MOLTO evaluation and assessment report*. Technical report, MOLTO project May 2013. `http://www.molto-project.eu/biblio/deliverable/d92-molto-evaluation-and-assessment-report`.

Snover, Matthew, Nitin Madnani, Bonnie 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 at the 12th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.

Tatsumi, Midori and Johann Roturier. 2010. Source Text Characteristics and Technical and Temporal Post-Editing Effort: What is Their Relationship? In *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC 10)*, Denver, USA. 43–51.

Tatsumi, Midori, Takako Aikawa, Kentaro Yamamoto and Hitoshi Isahara. 2012. How Good Is Crowd Post-Editing? Its Potential and Limitations. *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP 2012)*, San Diego, USA. 69–77.