Effects of Word Alignment Visualization on Post-Editing Quality & Speed[†]

Lane Schwartz lanes@illinois.edu
Dept. of Linguistics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Isabel Lacruz ilacruz@kent.edu
Tatyana Bystrova tbystro2@kent.edu
Dept. of Modern & Classical Language Studies, Kent State University, Kent, OH 44242, USA

Abstract

Phrase-based machine translation can be configured to produce alignment data that indicates which machine translated target language words correspond to which original source language words. In most prior work that examined the efficacy of post-editing machine translation, post-editors were presented with machine translations (and in most cases the original source language sentences) without also being presented with source-to-target alignment links. We select four news articles, and ask six Russian-English bilinguals and eleven Spanish-English bilinguals to post-edit English machine translation results, in some cases using alignments and in other cases without. We obtain human adequacy judgements of the post-edited sentences, and demonstrate that when machine translation quality is low, post-editing quality is consistently higher, by a statistically significant amount, when bilingual post-editors are presented with alignment data.

1 Introduction

Post-editing is the process whereby a human user corrects the output of a machine translation system. The use of basic post-editing tools by bilingual human translators has been shown to yield substantial increases in terms of productivity (Plitt and Masselot, 2010) as well as improvements in translation quality (Green et al., 2013) when compared to bilingual human translators working without assistance from machine translation and post-editing tools. More sophisticated interactive interfaces (Langlais et al., 2000; Barrachina et al., 2009; Koehn, 2009b; Denkowski and Lavie, 2012) may also provide benefit (Koehn, 2009a).

The question of how a post-editing interface should be configured and presented to users is a fundamentally interdisciplinary and empirical one. Issues of user interface design, human factors, translation studies, and machine translation quality are all likely relevant. Phrase-based machine translation can be configured to produce alignment data that indicates which machine translated target language words correspond to which original source language words. In most prior work that examined the efficacy of post-editing machine translation, post-editors were presented with machine translations (and in most cases the original source language sentences) without also being presented with source-to-target alignment links.

This work begins an attempt to answer two novel questions regarding post-editing interface design: To what extent, if at all, does the presentation of source-to-target word-level alignment links affect the quality or speed of post-editing? Is any such effect, if it exists, dependent on certain aspects of machine translation quality, or on the language pair?

[†]All code, scripts, data & analysis files for this paper are at `https://github.com/dowobeha/MT_Summit_2015`

(a) Russian-English adequacy evaluation guidelines

12	The post-edited translation is superior to the reference translation
10	The meaning of the Russian sentence is fully conveyed in the English translation
8	Most of the meaning of the Russian sentence is conveyed in the English translation
6	The English translation misunderstands the Russian sentence in a major way, or has many small mistakes
4	Very little information from the Russian sentence is conveyed in the English translation
2	The English translation makes no sense at all

(b) Spanish-English adequacy evaluation guidelines

10	The meaning of the Spanish sentence is fully conveyed in the English translation
8	Most of the meaning of the Spanish sentence is conveyed in the English translation
6	The English translation misunderstands the Spanish sentence in a major way, or has many small mistakes
4	Very little information from the Spanish sentence is conveyed in the English translation
2	The English translation makes no sense at all

Table 1: Adequacy evaluation guidelines for bilingual Russian-English human judges (Schwartz et al., 2014), and for bilingual Spanish-English human judges (Albrecht et al., 2009). Because no reference translation was available for Spanish-English, the 12 category is omitted.

To address these questions, we conduct a bilingual post-editing experiment (§2) where bilingual post-editors are presented with machine translation output of varying quality, with and without word-level alignment link visualization. In the first condition, we ask six Russian-English bilingual translation students to post-edit two Russian language news articles starting with relatively low quality English machine translation. In the second condition, we ask eleven Spanish-English bilingual translation students to post-edit two Spanish language news articles starting with relatively high quality English machine translation. We find (§3) that when machine translation quality is low, post-editing quality is consistently higher, by a statistically significant amount, when bilingual post-editors are presented with alignment data. We find no statistically significant effect when machine translation quality is high. We also found that for both Russian-English and Spanish-English the mean post-editing times were shorter for texts with alignment than for texts without alignment. These differences were not significant, but the difference for the Russian-English texts approached significance. Finally, in §4 we briefly survey the current state of post-editing research and situate this work within the context of related work in post-editing.

2 Methodology

We hypothesize that the presentation of word-level alignment links to human post-editors may affect the quality or speed of the resulting output, and that such effects may be dependent on the quality of the underlying machine translations. To test this hypothesis, we conduct a bilingual post-editing experiment where bilingual post-editors are presented with machine translation output of varying quality, with and without word-level alignment link visualization.

2.1 Bilingual Participants

2.1.1 Russian-English Bilingual Participants

There were six participants who served as post-editors in the Russian-English portion of this study, all of whom were paid for their time at the rate of $25 for each hour or part of an hour.

These participants were all English-Russian bilinguals. We designate these participants as PE1–PE6. Four of the six bilingual participants (PE2, PE3, PE4, & PE6) had Russian as their first language (L1) and were highly proficient in English as their second language (L2). The other two bilingual participants (PE1 & PE5) had English as their first language and were highly proficient in Russian as their second language. Three of the six bilingual participants were graduate students and three were undergraduate students; all were enrolled in a university Russian Translation program.

2.1.2 Spanish-English Bilingual Participants

There were eleven participants who served as post-editors in the Spanish-English portion of this study, all of whom were paid for their time at the rate of $25 for each hour or part of an hour. They were all Spanish-English bilinguals. We designate these participants as PE7–PE17. The first language (L1) of all eleven participants was English, and all eleven were highly proficient in Spanish as their second language (L2). These participants were students in a university Master of Spanish Translation program.

2.2 Source Language Data

2.2.1 Russian Data

For the Russian-English portion of this study, we selected as source texts a subset of the texts from the 2014 Workshop on Statistical Machine Translation (WMT-14) shared translation task (Bojar et al., 2014). Source texts were news articles covering world news events in late 2013. The first text was originally a Russian-language BBC news article covering Syrian chemical weapons. The second text was originally an English-language news article covering U.S. spying policy. We designate the former as Doc A and the latter as Doc B.

These two texts were each divided into segments (32 and 33 segments, respectively) that corresponded to sentences or stand-alone phrases (typically corresponding to news headlines, captions, or cutlines). Segments in Doc A varied in length from 3 to 35 words (mean length 17 words); segments in Doc B varied in length from 9 to 55 words (mean length 23 words).

Professional translations of Doc A into English and Doc B into Russian were commissioned as part of the WMT-14 shared translation task (Bojar et al., 2014). The Russian version of each text was translated automatically using Moses (Koehn et al., 2007) by Schwartz et al. (2014) as part of their WMT14 shared task submission. As a side effect of the phrase-based MT process, Moses can be configured to produce alignment links, indicating which target language words were produced from which source language words. To enable maximal comparability with the post-editing results of Schwartz et al. (2014), we make use of Russian-English machine translation results and alignments from that work here.

2.2.2 Spanish Data

Two Spanish source texts were selected. Both were extracts from a news article from a Spanish newspaper covering current world news events. The two texts were divided up into segments that corresponded to sentences or stand-alone phrases. The first text had 26 segments that varied in length from 4 to 24 words (mean length 15 words) and the second text had 25 segments that varied in length from 4 to 28 words (mean length 16 words). We designate the former as Doc C and the latter as Doc D. No reference translations exist for either Spanish text.

The Spanish source texts were translated automatically using Microsoft Bing Translator through its online developer API. Bing Translator, when accessed via the developer API, can be configured to return character-level alignment links from source characters to target characters, in addition to translated target language sentences. Our scripts derive word alignments from the character alignments returned by the Bing Translator API.

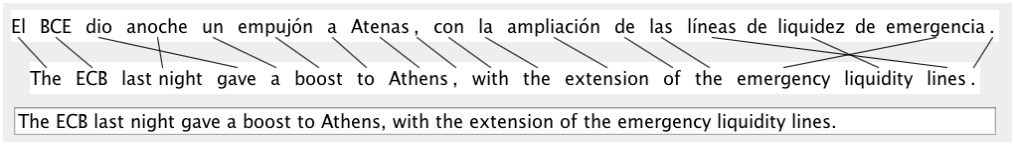


Figure 1: Post-editing interface, with alignment links displayed. Sentence shown is from Spanish-English Doc C.

2.3 Translation Quality

We hypothesize that any effects of word alignment visualization on post-editing may be dependent on the quality of the underlying machine translations displayed to the post-editors. Because we care about the adequacy of post-edited translations, we consider actual human judgements to be preferable to automated metrics such as BLEU (Papineni et al., 2002), which at best serve as a flawed proxy for human judgements. Instead, following Albrecht et al. (2009) and Schwartz et al. (2014), we therefore obtained human judgements of translation adequacy for the Russian-English and Spanish-English machine translations used in this study.

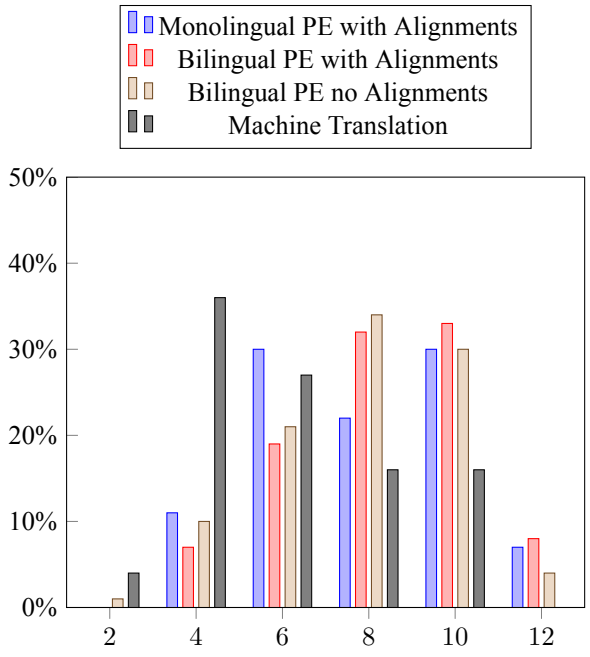
The Russian language news articles used in this study have corresponding reference translations. It is therefore possible (although given current machine translation quality, highly unlikely) that machine translation quality for any given segment could conceivably surpass the quality of the corresponding reference translation (if for example, the reference translator makes a mistake). For assessing the quality of the Russian-English machine translations, then, we follow the 12-point adequacy scale of Schwartz et al. (2014). This adequacy scale is shown in Table 1a on page 2; this scale ranges from a low of 2 (the English translation makes no sense at all) to a high of 12 (the translation is superior to the reference).

The Spanish language news articles used in this study lack corresponding reference translations. Thus, unlike the case of our Russian data, no matter how high the quality of machine translations, no Spanish-English machine translation segment could possibly receive a score of 12. For Spanish-English, we therefore follow the 10-point adequacy scale of Albrecht et al. (2009). This adequacy scale is shown in Table 1b on page 2; this scale is very similar to the former, but has a high of 10 (the meaning of the source sentence is fully conveyed in the English translation) instead of 12.

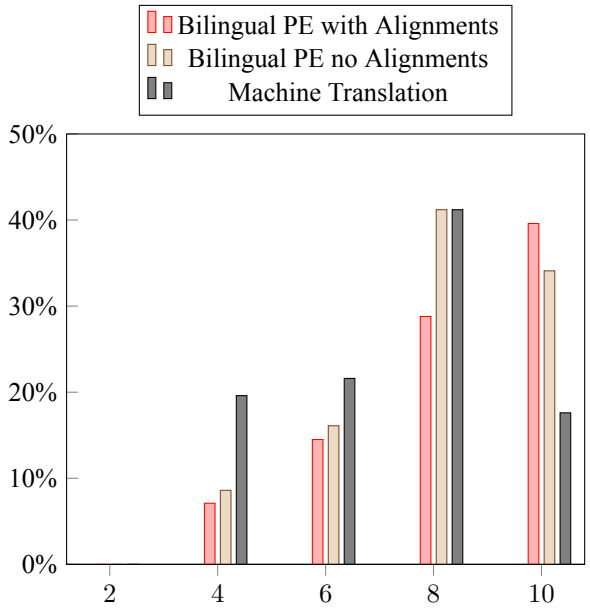
2.4 Post-Editing Interface

For this study, we developed a novel post-editing interface, based on the open source software used and released by Schwartz et al. (2014). Our software is written using Scala (Odersky, 2014), and is released as open source (see the software supplement that accompanies this work). This code constitutes a ground-up rewrite of the Java-based post-editing interface of Schwartz et al. (2014), written using a strict model-view-controller software design pattern to be easy for other researchers to use and extend.

Our post-editing interface can be seen in Figure 1 above. Each text was presented to post-editors in one of two variant modalities — word-level alignment links could either be visualized or left absent. In both variants, each source language segment was presented along with the corresponding machine translated English segment; a text field (initially populated with the machine translated segment) where the post-editor could make changes was also presented. In the first variant, the word-level alignment links produced by the machine translation decoder (Moses for Russian-English, Bing Translator for Spanish-English) were graphically displayed, linking source words to their corresponding machine translated target words. In the second variant, the word-level alignment links were omitted from the visualization interface.



(a) Russian-English



(b) Spanish-English

Figure 2: Percentage of segments judged to be in each adequacy category. For each language pair, we report percentages for raw (unedited) machine translation output, as well as output post-edited by a bilingual post-editor with access to alignments and without access to alignments. For Russian-English, we additionally report percentages for output post-edited by a monolingual post-editor (Schwartz et al., 2014) with access to alignments.

2.5 Procedure

Participants performed the test individually in an office setting and were instructed to minimally post-edit. Specifically, participants were asked to disregard issues of style and to focus on a) how well the translation conveyed the meaning of the source text, and b) the grammatical correctness of the translated segments. Participants sat in front of a computer that displayed the source texts divided up into segments (see Figure 1 on page 4). Directly below each source text segment, its machine translation was displayed. Below that was an active response area, where participants were asked to carry out the post-editing.

During initial data collection (the Russian-English condition), the only data collected was the final post-edited output and the overall time taken per text. Subsequently, we enhanced the post-editing software with additional logging functionality, enabling the software to record key-logging and mouse-logging data. For the subsequent Spanish-English condition, this enhanced software was utilized; for this condition millisecond-precision keyboard-event and mouse-event logs were recorded in addition to collecting final post-edited output and overall time taken per text.

We believe that scientific inquiry is at its strongest when experiments can be easily replicated, and when the raw and processed data from such experiments can be directly verified by reviewers, readers, and other experimenters. In that spirit, all data and code produced or used in this work are provided in the attached dataset and software supplements. This includes all logs, along with raw machine translation output, alignment data, post-edited output, adequacy judgements, post-editing software, and supplementary scripts.

2.5.1 Russian-English Participant Assignment

Each participant was instructed to edit one of the two texts using the interface where alignment links were shown, and to edit the other text using the interface where alignment links were omitted. Participants post-edited the two texts in one session lasting less than two hours, although there were no time limits set for the task. The experimenter was present in the room and manually recorded the times taken. Post-Editors 2, 4, and 6 were assigned to post-edit Doc A using the variant 1 interface that displayed alignments, and Doc B using the variant 2 interface that omitted alignments. Post-Editors 1, 3, and 5 were assigned Doc A using variant 2 and Doc B using variant 1.

2.5.2 Spanish-English Participant Assignment

Participants post-edited one text using the interface where alignment links were shown and the other text using the interface where alignment links were omitted. Participants post-edited the two texts in one session lasting less than one hour. Timings were recorded by a keylogger. Post-editors 7, 9, 11, 13, 15, and 17 edited Doc C using the interface that omitted the alignments and Doc D using the interface that displayed the alignments. Post-editors 8, 10, 12, 14, and 16 edited Doc C using the interface that displayed the alignments and Doc D using the interface that omitted the alignments.

3 Results

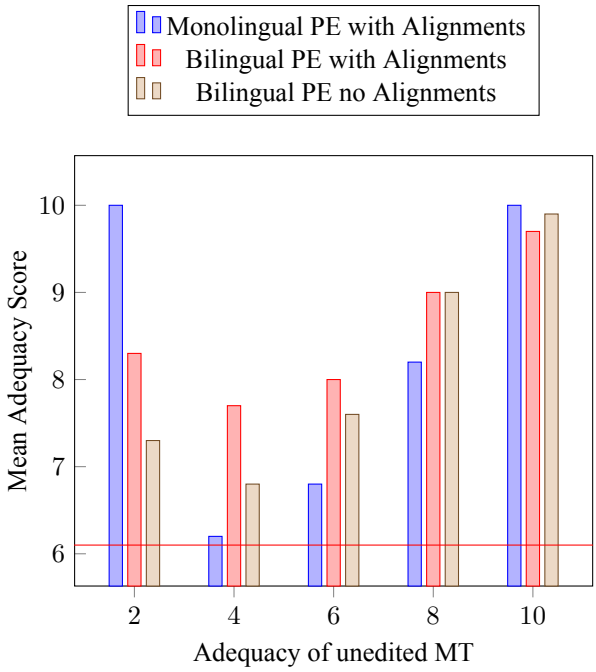
3.1 Rating Translation Adequacy

Following the methodology outlined in §2.3, all post-edited output as well as all machine translations were evaluated by bilingual human judges using the adequacy scales shown in Table 1.

3.1.1 Rating Adequacy of Russian-English

Following the adequacy guidelines from §2.3, an experienced English-Russian translator and grader rated all English output translations of the Russian-English post-edited segments. In addition, all English machine translations of the Russian documents were manually rated for

(a) Russian-English



(b) Spanish-English

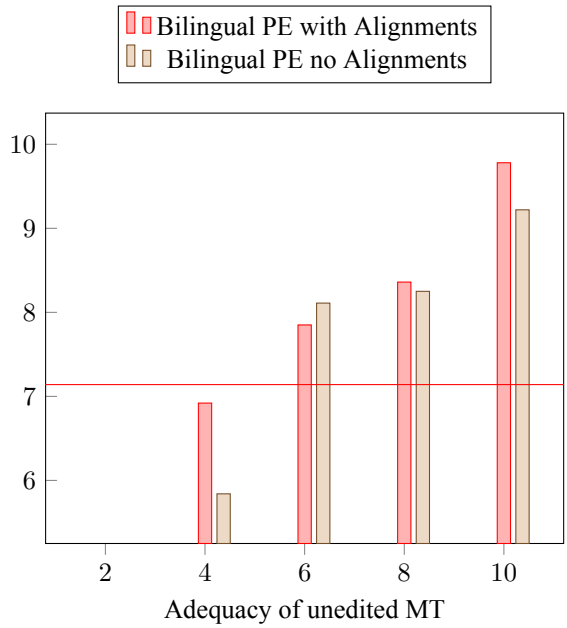


Figure 3: Mean adequacy score, categorized by the adequacy score of the unedited MT. The red horizontal line indicates the mean adequacy score (Russian-English: 6.1; Spanish-English: 7.1) of the unedited MT.

adequacy.

For each segment, the human rater was presented with a vertically-arranged list showing all variants of that segment. The first entry in each list was the segment in the source language (Russian). The source segment was followed by the reference translation in English. The subsequent eight entries were English translations of the source segment, presented in a randomized order. The English translations included the unedited machine translation output, as produced by Moses, a post-edited translation produced by a monolingual post-editor from Schwartz et al. (2014), and the six post-edited translations produced by the Russian-English bilingual post-editors in this study.

All Russian-English translations were rated using the translation adequacy scale in Table 1a on page 2, with possible ratings ranging from 2 (translation makes no sense) to 12 (translation is superior to the reference translation).

3.1.2 Rating Adequacy of Spanish-English

Following the adequacy guidelines from §2.3, an experienced Spanish-English translator and grader worked in cooperation with a second Spanish-English bilingual to rate all English output translations of the Spanish-English post-edited segments. In addition, all English machine translations of the Spanish documents were manually rated for adequacy. For each segment, the human raters were presented with a vertically-arranged list showing all variants of that segment. The first entry in each list was the segment in the source language (Spanish). The subsequent twelve entries were English translations of the source segment, presented in a randomized order. The English translations included the unedited machine translation output, as produced by Bing Translator, and the eleven post-edited translations produced by the Spanish-English bilingual post-editors in this study.

Unlike the Russian documents, no reference translation was available for the Spanish documents; for this reason (as described in §2.3), the top category (12) used in evaluating Russian-English segments was omitted from the Spanish-English evaluation. All Spanish-English translations were rated using the translation adequacy scale in Table 1b on page 2, with possible ratings ranging from 2 (translation makes no sense) to 10 (the meaning of the Spanish sentence is fully conveyed in the English translation). Unlike all other participants, participant PE17 consistently produced post-edited segments of lower adequacy than the corresponding raw MT output. This participant was therefore dropped from all analyses of the Spanish-English data.

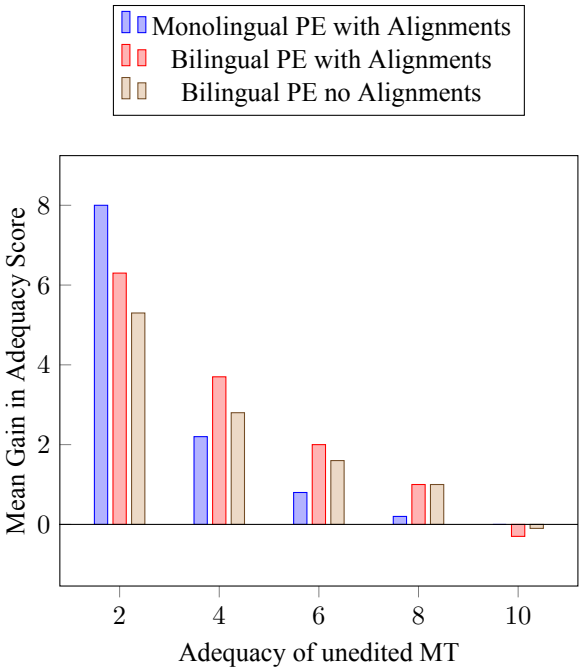
3.2 Adequacy Results

Figure 2 on page 5 presents the percentage of segments judged to be in each adequacy category. Mean adequacy scores for each experimental condition are presented in Figure 3 on the previous page. By subtracting the adequacy score of each machine translated segment, we obtain the adequacy gain obtained by post-editing; these values are presented in Figure 4 on the following page. Finally, Figure 5 on page 11 presents mean adequacy score by post-editor. We now analyze these results by experimental condition.

3.2.1 Machine Translation Adequacy

In Figure 2 on page 5, we observe that the Russian-English machine translation segments tend to be of lower quality (as measured by adequacy), while the Spanish-English machine translation segments tend to be of higher quality. Two-thirds of Russian-English machine translation segments are judged to have major errors (ratings 2-6), while one-third are rated as mostly or completely correct (8-12). Contrast this with the Spanish-English machine translations segments; a minority (about two-fifths) are judged to have major errors (ratings 2-6), while the majority (about three-fifths) are rated as mostly or completely correct (8-10).

(a) Russian-English



(b) Spanish-English

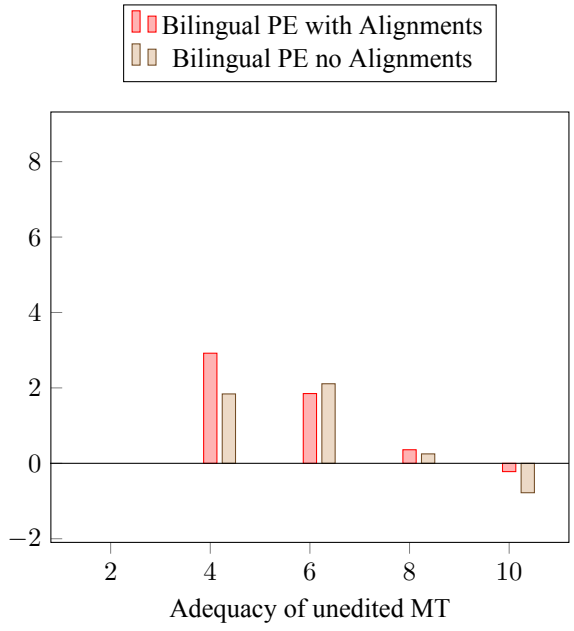


Figure 4: Mean gain in adequacy score over unedited MT, categorized by the adequacy score of the unedited MT.

3.2.2 Russian-English Adequacy

The mean adequacy score when bilingual participants were presented with alignments was 8.35. When alignments were omitted from the post-editing tool, the mean adequacy score was 7.85. A Wilcoxon signed-rank test (Wilcoxon, 1945) showed that when participants were presented with alignments the ratings of their translations were significantly higher than when participants post-edited without access to alignments ($N = 6$, $Z = -2.207$, $p = 0.027$).

3.2.3 Spanish-English Adequacy

The mean adequacy score when Spanish bilingual participants were presented with alignments was 8.22. When alignments were omitted from the post-editing tool, the mean adequacy score was 8.02. These means are not significantly different.

3.3 Timing Results

3.3.1 Russian-English Timing

The times taken by the Russian-English post-editors to post-edit each text were recorded manually to the nearest minute. Participants post-edited each text without a break. They took a break of approximately five minutes between texts. The mean times were 33 minutes for texts with alignment and 40 minutes for texts without alignment. This difference approached significance ($p = .082$).

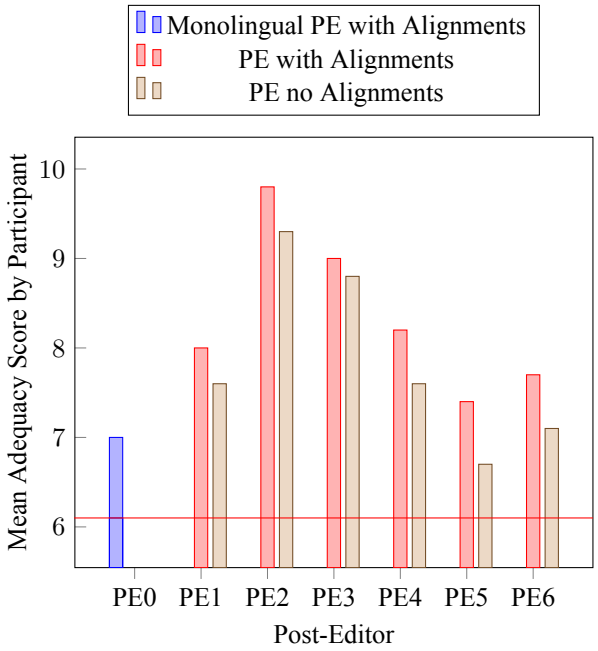
3.3.2 Spanish-English Timing

The times taken by the Spanish-English post-editors to post-edit each text were recorded by the keylogger to nearest millisecond. Participants post-edited each text without a break. They took a break of approximately five minutes between texts. The mean times were 21 minutes and 5 seconds for texts with alignment and 22 minutes and 5 seconds for texts without alignment. An independent samples t-test showed that these times are not significantly different from each other ($t(18) = .295$, $p = .77$.) Participants were not given time limitations, so timing data must be interpreted with caution. Note, however, that the mean time with alignment is numerically shorter than the mean time without alignment. The shorter editing times for Spanish-English may in part be explained by the shorter length of these documents.

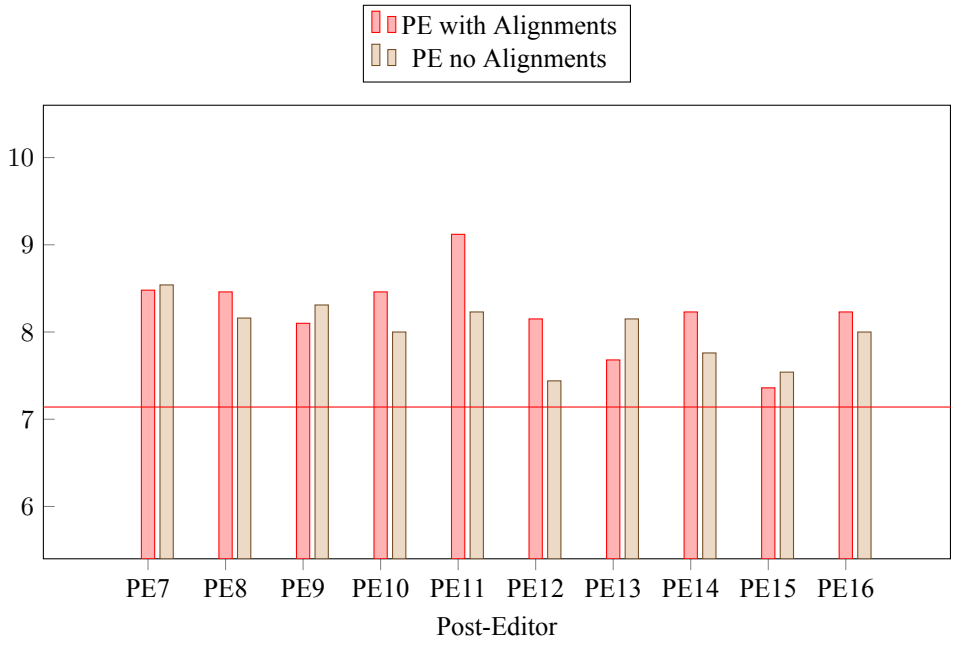
4 Analysis and Related Work

Our results suggest that when machine translation quality is poor (2–4), bilingual post-editors may produce higher quality translations when presented with bilingual alignment links between source words and machine-translated target words. Alternatively, when machine translation quality is high (8–12), no effect is seen by presenting alignment visualizations. We explain this by hypothesizing that word alignment visualization may enable post-editors to better recover from certain types of translation errors produced by MT systems; when MT quality is high enough that such errors are absent, word alignment visualization may no longer play a restorative role.

We examine the effect that alignment link visualization has on each bilingual post-editor in Figure 5 on the next page. In the Russian-English condition, where overall MT quality is poor, we observe that post-editing quality varies widely between post-editors (with PE2 and PE3 performing best). For all six bilingual post-editors, we observe higher mean adequacy scores when alignment links were presented than when they were omitted from the post-editing tool. We also note that when alignment links were absent, one bilingual post-editor (PE5) performed worse than the monolingual post-editor (PE0) from Schwartz et al. (2014). On the other hand, in the Spanish-English condition, where overall MT quality is good, we observe relatively little variation in quality between the ten post-editors. When compared to the unedited machine trans-



(a) Russian-English



(b) Spanish-English

Figure 5: Mean adequacy score per post-editor. The red horizontal line indicates the mean adequacy score (Russian-English: 6.1; Spanish-English: 7.1) of the unedited MT.

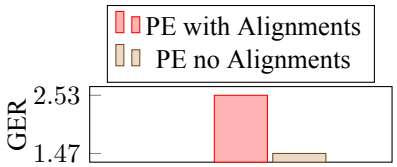


Figure 6: Mean adequacy gain to effort ratio (GER) values for segments post-edited with alignment and without alignment. Effort is measured by pause to word ratio (PWR).

lations, post-editing resulted in improved mean adequacy for all post-editors, both bilingual and monolingual.

Our results also suggest that post-editing time tends to be reduced for texts with alignment. These numerical reductions were consistent across all the Russian-English participants with the exception of PE3, but they were not consistent for the Spanish-English participants.

We hypothesize that texts with alignment are less cognitively demanding to process, and so less effortful to post-edit than texts without alignment. If this is the case, shorter post-editing times for texts with alignment are consistent with previous findings by Koponen et al. (2012), who found that per word post-editing times were shorter for segments that were less cognitively demanding because of the linguistic structure. Related work on cognitive effort in post-editing (Lacruz et al., 2014; Lacruz and Shreve, 2014) has also shown decreased densities of short pauses when less cognitively demanding segments are post-edited.

The keystroke logging data gathered for Spanish-English post-editors allowed the computation of Pause to Word Ratio (PWR). For each segment, PWR is the ratio of the number of pauses exceeding 300ms to the number of words in the MT segment; it is a measure of cognitive effort in post-editing (Lacruz and Shreve, 2014). Higher PWR corresponds to higher cognitive effort. Contrary to expectation, the mean PWR for Spanish-English post-editors was slightly higher for the segments with alignment (0.70) than for those without alignment (0.63). However, the numerical difference was not significant.

It is possible that the effect of alignment on PWR was masked by the fact that the adequacy of the Spanish-English MT segments was generally high. Since our prediction is that alignment should both increase post-editing adequacy and reduce post-editing effort, the Gain to Effort Ratio, GER = (PE Rating – MT Rating)/PWR is a promising metric to investigate. We hypothesize that GER is higher for segments with alignment than for segments without alignment.

Figure 6 above shows GER for Spanish-English. GER for segments with alignment was 2.53, and GER for segments without alignment was 1.47. Our prediction was confirmed: a paired samples t-test showed that GER was higher for segments with alignment ($t(9) = 2.49$, $p = .034$.). This result suggests that GER may be a robust metric for measuring the effects of alignment on post-editing. It would be interesting to conduct further studies involving language pairs different from Spanish-English where the adequacy of machine translations may be lower.

5 Conclusion

In this work, we observe that when machine translation quality is poor, bilingual post-editors may produce higher quality translations when presented with bilingual alignment links between source words and machine-translated target words. We explain this by hypothesizing that word alignment visualization may enable post-editors to better recover from certain types of translation errors produced by MT systems; when MT quality is high enough that such errors are absent, word alignment visualization may no longer play a restorative role. The timing results we observe, while not statistically significant, appear to be consistent with prior work that found per word post-editing times to be shorter for segments that were less cognitively demanding.

References

Albrecht, J. S., Hwa, R., and Marai, G. E. (2009). Correcting automatic translations through collaborations between MT and monolingual target-language users. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 60–68, Athens, Greece.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., and Vilar, J.-M. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA.

Denkowski, M. and Lavie, A. (2012). TransCenter: Web-based translation research suite. In *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice Demo Session*.

Green, S., Heer, J., and Manning, C. D. (2013). The efficacy of human post-editing for language translation. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, pages 439–448, Paris, France.

Koehn, P. (2009a). A process study of computer aided translation. *Machine Translation*, 23(4):241–263.

Koehn, P. (2009b). A web-based interactive computer aided translation tool. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 17–20, Suntec, Singapore.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL '07) Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Koponen, M., Aziz, W., Ramos, L., and Specia, L. (2012). Post-editing time as a measure of cognitive effort. In *AMTA 2012 Workshop on Post-Editing Technology and Practice*, WPTP, pages 11–20, San Diego, USA.

Lacruz, I., Denkowski, M., and Lavie, A. (2014). Cognitive demand and cognitive effort in post-editing. In *Proceedings of the AMTA 2014 Workshop on Post-Editing Technology and Practice*, pages 73–84.

Lacruz, I. and Shreve, G. (2014). Pauses and cognitive effort in post-editing. In O'Brien, S., Balling, L. W., Carl, M., Simard, M., and Specia, L., editors, *Post-Editing: Processes, technology and applications*, pages 246–272.

Langlais, P., Foster, G., and Lapalme, G. (2000). TransType: A computer-aided translation typing system. In *Proceedings of the ANLP/NAACL 2000 Workshop on Embedded Machine Translation Systems*, pages 46–51, Seattle, Washington.

Odersky, M. (2014). The Scala language specification, version 2.9. Technical report, EPFL Programming Methods Laboratory, Lausanne, Switzerland.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*, pages 311–318, Philadelphia, Pennsylvania.

Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Schwartz, L., Anderson, T., Gwinnup, J., and Young, K. M. (2014). Machine translation and monolingual postediting: The AFRL WMT-14 system. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 186–194, Baltimore, Maryland. Association for Computational Linguistics.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.