Translation Quality and Effort: Options versus Post-editing

Donald Sturgeon[*]
Fairbank Center for Chinese Studies
Harvard University
djs@dsturgeon.net

John S. Y. Lee
Department of Linguistics and Translation
City University of Hong Kong
jsylee@cityu.edu.hk

Abstract

Past research has shown that various types of computer assistance can reduce translation effort and improve translation quality over manual translation. This paper directly compares two common assistance types – selection from lists of translation options, and post-editing of machine translation (MT) output produced by Google Translate – across two significantly different subject domains for Chinese-to-English translation. In terms of translation effort, we found that the use of options can require less technical effort than MT post-editing for a domain that yields lower-quality MT output. In terms of translation quality, our analysis suggests that individual preferences play a more decisive role than assistance type: a translator tends to consistently work better either with options or with MT post-editing, regardless of domain and of their translation ability.

1 Introduction

State-of-the-art computer-assisted translation (CAT) systems offer many types of assistance to the human translator. Most studies have focused on investigating whether such assistance — including translation memory, Machine Translation (MT) output, and word and phrase translation options — results in higher productivity and better quality when compared with unassisted translation (Plitt & Masselot, 2010; Zhechev, 2012; Green et al., 2013; Aranberri et al., 2014; Gaspari et al., 2014). Less attention, however, has been devoted to comparing the relative merits of these assistance types. This paper presents a direct comparison between two common types, namely, selection from translation options, and post-editing of MT output, with Google Translate as the MT system. We analyze these two assistance types along two dimensions:

Translation effort. Translation effort can be measured in terms of time or the amount of editing. Previous research has found between-translator variance of the number of post-editing operations to be lower than that of post-editing time (Tatsumi and Roturier, 2010; Koponen, 2012; Koponen et al., 2012). Therefore, we will hold temporal effort as constant, and instead measure technical effort (Krings, 2001), by explicitly measuring the amount of editing and the number of clicks needed for selecting options. We investigate which assistance type requires less effort, across two domains that differ in terms of MT output quality.

Translation quality. We measure whether either of these assistance types results in better human translations. Past studies have suggested that individual work style is an important factor (e.g., Koehn and Germann, 2014); this study provides further evidence by analyzing a number of other possible factors, including the effect of different domains and translator abilities.

2 Previous work

As publicly available MT systems continue to improve, human translators increasingly adopt post-editing of MT output to boost their productivity. Naturally, the quality of the output is a major factor that determines its benefits. It has been shown that better MT systems generally yield greater productivity gains (Koehn and Germann, 2014). Even state-of-the-art MT systems, however, tend to produce lower-quality output for source sentences from more specialized domains, because of mismatched data. For these domains, the available bilingual data is often insufficient to train a statistical MT system. An alternative is to infer word and phrase alignments from the limited data available,

[*] The first author conducted this research at the Department of Linguistics and Translation at City University of Hong Kong.

to be offered as translation options to the human translators. A research question that remains under-investigated is whether the use of options or MT output post-editing is more suitable for domains with varying levels of MT output quality.

Two previous studies have evaluated both the use of options and MT post-editing,[1] but on only a single domain. One involved monolingual translators with no knowledge of the source language (Koehn, 2010). They were asked to translate news stories from Arabic and Chinese to English, aided by post-editing and translation options. The study found no significant difference between the two assistance types for Arabic but better performance with options for Chinese. The other study was concerned with French-to-English translation on the news domain (Koehn, 2009b). Each of three kinds of assistance – prediction of the next word/phrase, options, and MT post-editing – improved translation productivity and quality overall. Most relevant for our study, MT post-editing outperformed options, saving 1 second per word and achieving a 4% higher correctness rate (Koehn, 2009b:p.250, Table 2).

This paper further compares these two assistance types in terms of a number of other factors. One factor is the subject domain. We consider whether two domains, one resource-rich and the other resource-poor, may favor different assistance types. Another factor is individual preference for specific assistance types. In the aforementioned study, half of the subjects achieved higher rates of correct translations with options, and the other half the opposite (Koehn, 2009b:p.250, Table 2). Using a larger pool of subjects, we investigate correlations between translation quality and other variables, namely, different domains and translator abilities.

3 Experimental setup

3.1 Domains

We chose two contrasting domains on which to conduct our experiments. The first is a resource-rich domain with many commercial MT systems trained with similar bilingual data; the second is resource-poor with limited samples of bilingual sentences, and would be considered out-of-domain for most commercial MT systems.

- ***“Multi UN” domain***. This domain consists of a corpus of United Nations documents published between January 2000 and March 2010 (Eisele and Chen, 2010). Among the largest parallel corpora available for Chinese and English (Tian et. al, 2014), the corpus has a total of 9.6 million aligned sentences.
- ***“Literary” domain***. This domain is derived from the first 51 chapters of the Chinese classic novel, *Romance of the Three Kingdoms*, and an English translation by C.H. Brewitt-Taylor.[2] These chapters contained 1563 paragraph alignments and 249390 characters.

3.2 Translation assistance types

For each domain, we compared two translation assistance types:

- ***Translation options***. The user first selects translation options for phrases in the source sentence, and then further edits the resulting target sentence (see next section for a description of the interface). For each of the two domains we compiled a phrase table to store these options. Both tables contain bilingual dictionary data from CEDICT (cc-cedict.org) and from the Chinese Text Project (ctext.org). The table for the Multi UN domain is further enriched with word and phrase correspondences extracted from the corpus of UN documents (Eisele and Chen, 2010); the table for the Literary domain, with those from the aforementioned bilingual data from the *Romance of the Three Kingdoms*. While the UN corpus provides sentence alignments,[3] we aligned the Chinese and English texts in the Literary domain using Microsoft’s “Bilingual Sentence Aligner” tool,[4] followed by manual review to exclude false matches. We then used Giza++

[1] Some other studies have focused on comparing post-editing from MT output and from sentences that are exact or fuzzy matches with entries in the phrase table (Guerberof, 2009; Laubli et al., 2013).

[2] http://ctext.org/sanguo-yanyi

[3] Sentences with over 40 Chinese characters were excluded.

[4] The aligner implements the algorithm described by Moore (2002). See http://research.microsoft.com/en-us/downloads/aafd5dcf-4dcc-49b2-8a22-f7055113e656/

(Och and Ney, 2003) to extract alignments between Chinese and English and used these to construct our phrase table.

- ***MT output post-editing***. The user post-edits the translation produced by a fully automatic MT system; in our case, the Google Translate system (translate.google.com). An alternative approach would have been to train statistical machine translation models using the respective bilingual datasets described in the previous section. We trained such a model for the Multi UN domain using Moses MT (Koehn et al., 2007); the resultant model achieved a TER of 56.9 on the test set (see Section 3.4) when measured against the reference translation, which was on a par with the TER of 58.3 achieved by Google Translate. For the Literary domain, however, it would be impractical to train such a model, given the limited amount of data available. A possible mitigation is to attempt domain adaptation (e.g., Song et al., 2012), but this method would introduce a possibly confounding variable. Instead, for more straightforward comparison, we made use of Google Translate, a general-purpose and widely used MT system, on both domains.

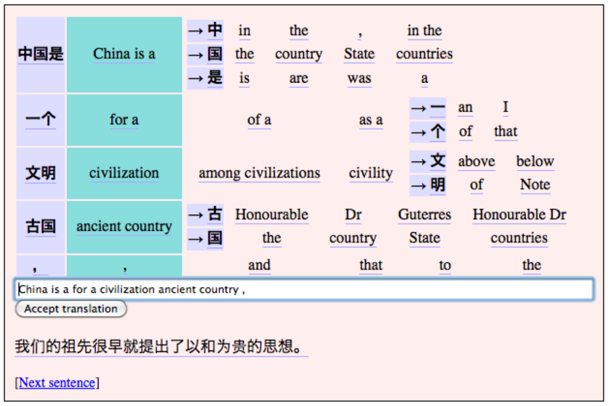
3.3 Interface

Figure 1 shows the interface used in our experiments for displaying translation options. The user is presented with a list of translation options for phrases in the source sentence, in decreasing order of the frequency with which they occur in the training data.[5] Such matches may be strings of any length, and are chosen using forward maximal matching against the phrase table. For each matched phrase, the translation with the highest frequency is pre-selected by default; for example, "China is a" is pre-selected on the first row in Figure 1. The user may substitute one of the proposed alternatives in place of the pre-selected option. Selections are made as an on-off toggle with a maximum of one selection per row, and a null selection can be made by clicking any selected translation to deselect it. Clicking on an item immediately substitutes it into the appropriate location in the proposed translation string. For example, in Figure 2, the English translation "contraction" has been selected over the default "austerity" on the fifth row; and the null selection has been made on the second and seventh rows. Once all such choices have been made, the user edits the string composed of the choices made for each phrase.

Additional features commonly included in production CAT systems, such as the ability to add new entries to the phrase table and to dynamically update frequencies of phrase table items were disabled during our experiment to avoid complicating our results with factors that would likely vary depending upon the amount of text reuse within each passage in the test set.

To make the translation procedure as similar as possible in both MT post-editing and options cases, the text for the MT post-editing evaluations was fetched beforehand from Google Translate and imported into the same environment used for the options experiments, where it was presented for post-editing in the same manner as for the options case, but without displaying any matches from a phrase table.

In what follows, we use "default selections" to refer to the initial selections as presented to the user (and their concatenation), "user selections" for the actual combination of selections ultimately chosen by the user (and the corresponding string), and "final translation" for the finished translation created by post-editing of the user selections.



中国是	China is a	→中 in the , in the →国 the country State countries →是 is are was a			
一个	for a	of a	as a	→一 an I →个 of that	
文明	civilization	among civilizations	civility	→文 above below →明 of Note	
古国	ancient country	→古 Honourable Dr Guterres Honourable Dr →国 the country State countries			
,	,	and	that	to	the

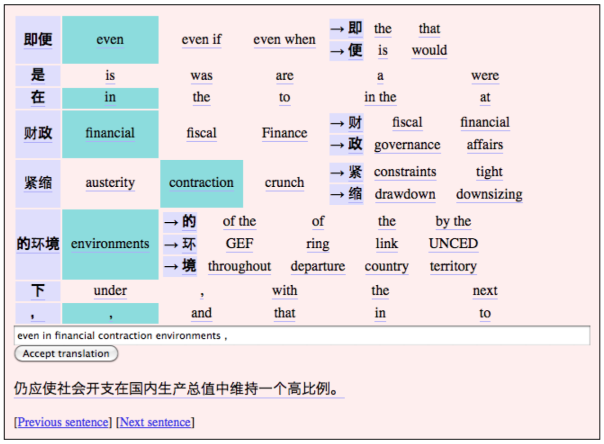
China is a for a civilization ancient country ,

我们的祖先很早就提出了以和为贵的思想。

[Next sentence]

Figure 1. The interface for displaying translation options, shown with default selections highlighted, on a Chinese-to-English translation task.

[5] We use the word "phrase" to refer to an arbitrary string of terms rather than any particular linguistic construction.



即便	even	even if	even when	→ 即 → 便	the is	that would
是	is	was	are	a	were	
在	in	the	to	in the	at	
财政	financial	fiscal	Finance	→ 财 → 政	fiscal governance	financial affairs
紧缩	austerity	contraction	crunch	→ 紧 → 缩	constraints drawdown	tight downsizing
的环境	environments	→ 的 → 环 → 境	of the GEF throughout	of ring departure	the link country	by the UNCED territory
下	under	,	with	the	next	
,	,	and	that	in	to	

even in financial contraction environments ,

仍应使社会开支在国内生产总值中维持一个高比例。

[Previous sentence] [Next sentence]

Figure 2. The interface for displaying translation options, with two null selections (rows 2 and 7) and one alternative selection (row 5). This set of selections corresponds to a total of three clicks, the minimum number required to produce this set of selections.

3.4 Procedure

Twenty-four Masters-level students in the Translation Department, at a university in which English is the primary language of teaching, participated in this study.[6] They were all native speakers of Mandarin and highly competent in English.

The students performed translations in four 45-minute sessions: Multi UN post-editing, Multi UN options, Literary post-editing, and Literary options. In each session, they translated continuous passages of Chinese text from the appropriate domain into English using the interface describe above. For the post-editing sessions, the output of Google Translate was provided in the same interface, with the options table hidden. The students were not allowed to use any additional translation aids, with the exception of a single specified online dictionary (cdict.net). The translation had to be completed in the allotted time; thus, we held translation time as a constant, and measured differences in translation effort and quality in the four sessions.

Our test sets contained data similar to but excluded from the data used to create the phrase table. For the Literary domain, this consisted of sections of text from chapters 52 onwards; for the Multi UN domain, we selected document fragments immediately post-dating the latest documents in the corpus.[7] For each domain, we used the same test set for both assistance types, while ensuring that each student translated a different source text using options to the text they post-edited from MT. We asked two different students to translate the same passage under each condition.

After the translations had been completed, each student was asked to evaluate the quality of the translations of four other students, two of whom had translated the same passage using options and two of whom had translated this same passage by post-editing MT output. These four translations were presented in a differing random order each time so that the evaluators had no way of knowing which translator or which assistance type corresponded to any given part of the translation. Students were asked to give a holistic rating from 1 to 5 of the quality of each translation. There was substantial agreement among the annotators. 43% of quality annotations differed by 1 or less between the two annotators who evaluated the same translation, and 87% differed by 2 or less.

4 Experimental results

We first discuss our results on the amount of translation effort required (Section 4.1), and then the factors impacting translation quality (Section 4.2).

4.1 Translation effort

Table 1 reports the average technical translation effort for each of the four sessions. For MT post-editing (“PE_{MT}”), we measure the effort using Translation Edit Rate (TER) (Snover et. al, 2006). This metric, used also in a number of previous studies (e.g., Koponen 2012; Koponen 2013; Koehn and Germann 2014), reflects the number of edit operations performed per 100 words of the final translation.

For the use of options, we measure two kinds of effort (“Options+PE_{user}”). The first kind is the effort for selecting options (“Options”). Rather than the raw number of clicks, we report instead the number of clicks (i.e. options chosen) per 100 words of the final translation, to facilitate a more

[6] Although a total of 25 students participated, one student failed to fully observe the rules of the experiment, and so this student’s data is excluded from the data presented in this paper.

[7] Document fragments from April 2010 available through http://documents.un.org/, the UN documents site from which the Multi UN corpus was extracted.

natural comparison with the TER figures. We considered only the final user selection – i.e. for a given word, clicking the second suggested translation, and then clicking the third suggestion thereby replacing it, would count as one click not two. The second kind is the effort for post-editing the user selections into the final translation (“PE_{user}”), again reported as TER. As a baseline, we also include the hypothetical TER for post-editing the default selections into the final translation without using options (“$PE_{default}$”).

As shown in Table 1, for every 100 words in the Literary domain, the use of options reduced the number of edits (TER) from 65.1 to 48.8, at the cost of 25.7 clicks. The editing effort compares favorably to MT post-editing, whose relatively high TER of 61.0 reflects the poor quality of MT output in the face of out-of-domain sentences. Assuming that less effort is needed to click than to edit a word, it can be argued that the use of options requires less technical effort for this domain.

In the Multi UN domain, the use of options again reduced the TER, from 47.4 to 36.1, at the cost of 12.0 clicks. The TER for MT post-editing, however, was even lower, at 27.3. This result suggests that for a resource-rich domain with MT systems trained with matching data, the higher quality of the MT output outweighs the reduction in effort brought by the options.

The number of clicks per 100 words is notably higher for the Literary domain than the Multi UN. This can be explained by modern Chinese having longer average word length than literary Chinese. Furthermore, UN documents tend to contain repeated technical terms and phrases that are identified and matched as single phrases when they reoccur.

One factor that could potentially influence our results is the degree to which the translator selects the option in the “optimal” manner. As observed by Koehn (2009b), when the translator saw an option that was suitable, he or she might have simply typed it in, rather than using the clicking mechanism to insert. To investigate whether this was a common phenomenon, we needed to compare actual user selection of options to the hypothetically “ideal” selections, given the final translation. To do this, we calculated the optimal set of selections that would have resulted in the closest string to that user’s final translation, as measured by TER. On the Literary domain, the selections chosen by several translators had the same TER as the optimal selections, and average user selection performance was within 10% of ideal performance.[8] On the Multi UN domain, only one translator’s selections scored the same as the ideal selections, and average user performance was within 20% of ideal. These figures confirm that while translators may have been influenced by options that they did not select, they only rarely failed to select options that would have reduced the technical effort required for their final translation.

Domain	Assistance Type	Clicks	TER
Multi UN	$PE_{default}$	0	47.4
	Options+PE_{user}	12.0	36.1
	PE_{MT}	0	27.3
Literary	$PE_{default}$	0	65.1
	Options+PE_{user}	25.7	48.8
	PE_{MT}	0	61.0

Table 1. Technical translation effort for using MT post-editing (“PE_{MT}”) and options (“Options+PE_{user}”), and the baseline of post-editing directly from default selections without using options (“$PE_{default}$”). The effort is expressed by the number of clicks per 100 words for selecting options, and by the TER for post-editing from Google Translate output (“PE_{MT}”) or from the user selections (“PE_{user}”).

4.2 Translation quality

Table 2 shows the average translation quality for different combinations of domains and assistance types. Despite their different levels of translation effort, all four yielded similar average quality scores, with MT post-editing on the Multi UN domain scoring slightly higher. This is likely explained in part by the fact that Google Translate uses UN documents as training data, and thus performs particularly well on material from this domain.[9] These averages, however, mask some

[8] That is, on average the amount of post-editing work that a user could have avoided if he or she had chosen precisely those options that would minimize such work was less than 10% of his or her total post-editing work, as measured by the TER metric.

[9] http://www.reuters.com/article/2007/03/28/us-google-translate-idUSN1921881520070328

significant variations, to which we now turn our attention.

Domain	Assistance Type	Quality score
Multi UN	PE_{MT}	4.25
	$Options+PE_{user}$	4.06
Literary	PE_{MT}	4.06
	$Options+PE_{user}$	4.09

Table 2. Average translation quality by assistance type and domain as measured by manual evaluation.

Options vs. post-editing. Of the 24 students, 10 scored higher in both domains when using options, 11 scored higher in both domains using post-editing; only three scored higher on the Literary domain using options but scored lower with it on the UN domain. There was thus a strong correlation between difference in options-based quality and post-edit-based quality in the two domains (Pearson correlation coefficient: r=0.84, p<0.01). Figure 3 illustrates this correlation as a graph; with the exception of the three data points in the lower-right quadrant, x-y pairs are always either both positive or both negative. In other words, the options consistently helped some students to create higher quality translations, while other students consistently produced higher quality translations by post-editing, even for two domains with significant differences in MT output quality.

We can thus divide our translators into two main groups: those who on both domains improved their translation quality with options (which we term the "Options+" group), and those who showed improved quality with MT post-editing ("Options-"). Table 3 shows the consistent gap in average quality score between options and MT post-editing for these two groups.

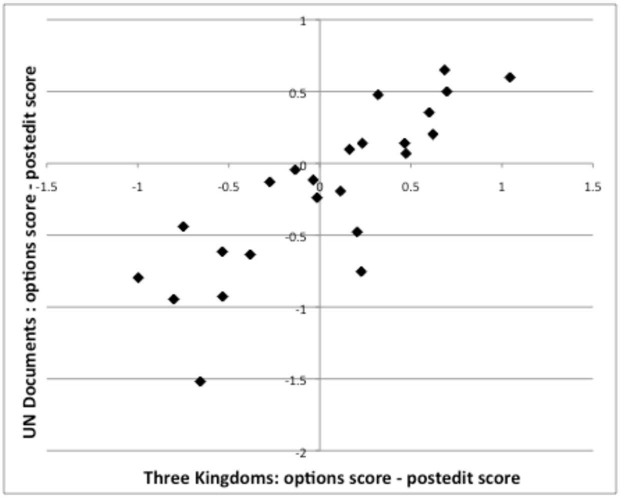


Figure 3. Improvement in manual quality assessment score using options on the Multi UN domain vs. improvement using options on the Literary domain.

Group	Domain	Assistance Type	Quality Score
Options+	Multi UN	PE_{MT}	3.93
		$Options+PE_{user}$	***4.25***
	Literary	PE_{MT}	3.83
		$Options+PE_{user}$	***4.36***
Options-	Multi UN	PE_{MT}	***4.45***
		$Options+PE_{user}$	3.87
	Literary	PE_{MT}	***4.22***
		$Options+PE_{user}$	3.76

Table 3. Translation quality as measured by manual evaluation, grouped by those whose translation quality increased with options (+) and those whose quality decreased (-).

Assistance uptake. We next investigated whether members of the Options- group were unable or unwilling to make use of the options, similar to the "refuseniks" identified by Koehn (2009b). As shown in Table 4, the average number of clicks per 100 words among members of the Options- group was somewhat below those of Options+ for both domains (22.6 vs 26.2 for Literary, and 11.1 vs 12.2 for Multi UN); but these figures are still broadly comparable, indicating that both groups did make use of options. Additionally, both the highest and lowest numbers of clicks per 100 words (2.0 and 47.2 respectively) for a document in the Literary domain occurred in the Options-group, strongly suggesting that other factors beyond whether or not users made use of options affected the degree to which they benefited from their presence. There was however a clear correlation (r=0.69, p<0.01) between how often an individual user clicked on one domain and on the other, indicating that user preference is a more important factor than domain in determining assistance uptake.

Domain	**Group**	**Clicks per 100 words**
Multi UN	Options+	12.2
	Options-	11.1
Literary	Options+	26.2
	Options-	22.6

Table 4. The average number of clicks (i.e., option selections) per 100 words, compared between the Options+ and Options- groups and across domains.

Translator ability. Finally, we considered the possibility that only more advanced translators benefited from the options, or vice versa. The average per-translator per-domain quality scores ranged from 3.3 to 4.9, confirming the existence of variation in individual translator ability. Looking at the absolute quality scores for the Options+ and Options- groups, we found the same average score of 4.1 for both groups when averaged over all four of each individual's translations. The strongest and weakest performing individuals were both in the Options+ group, demonstrating that both strong and weak translators can and do benefit from options, though not all do. These results suggest that translator ability does not determine whether or not options are beneficial.

In summary, regardless of the differences in domain and the quality of MT output, and regardless of their translation competence, some translators consistently produced better translations with MT post-editing than with options, even though they made full use of options; others showed the opposite tendency, again consistently so.

5 Conclusions

This paper has presented a study on the use of word and phrase translation options and MT post-editing. We compared a resource-rich domain that benefits from an MT system with matching training data, and a resource-poor one that yields lower-quality MT output. We found that the use of options required less technical effort than MT post-editing for the latter domain, but not for the former. In terms of translation quality, however, we found that individual translators exhibited consistent preferences for either options or MT post-editing across two domains: those whose translations improved when using options as compared with MT post-editing benefitted more from this type of assistance regardless of domain, and likewise those who did better without it again did so without respect to domain. Furthermore, we found that improvement in translation quality was not simply a function of translation ability, nor was it merely a matter of whether or not translators engaged with options selecting functionality of the CAT system. We therefore echo Koehn (2009b)'s suggestion that more study is needed into the cognitive processes of translation and how these may explain these different outcomes.

Acknowledgements

The work described in this paper was supported by a Teaching Development Grant from City University of Hong Kong (Project No. 6000489).

References

Nora Aranberri, Gorka Labaka, Arantza Diaz de Ilharraza, and Kepa Sarasola. 2014. Comparison of post-editing productivity between professional translators and lay users. *Proc. AMTA Workshop on Post-Editing Technology and Practice*.

Andreas Eisele, Yu Chen. 2010. MultiUN: A Multilingual Corpus from United Nation Documents. *Proc. Seventh conference on International Language Resources and Evaluation*, p. 2868-2872.

Federico Gaspari, Antonio Toral, Sudip Kumar Naskar, Declan Groves, and Andy Way. 2014. Perception vs. Reality: Measuring Machine Translation Post-editing Productivity. *Proc. Third Workshop on Post-Editing Technology and Practice*.

Spence Green, Jeffrey Heer, and Christopher D. Manning. 2013. The Efficacy of Human Post-Editing for Language Translation. *Proc. ACM Human Factors in Computing Systems (CHI).*

A. Guerberof. 2009. Productivity and quality in MT post-editing. 2009. *Proc. MT Summit Workshop on New Tools for Translators.*

Chung-chi Huang, Mei-hua Chen, Ping-che Yang and Jason S. Chang. 2013. A Computer-Assisted Translation and Writing System. *ACM Transactions on*

Asian Language Information Processing 12(4): Article 15.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation, *Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session, Prague, Czech Republic.

Philipp Koehn. 2009a. A web-based interactive computer aided translation tool. *Proc. ACL-IJCNLP 2009 Software Demonstrations*, p. 17-20.

Philipp Koehn. 2009b. A process study of computer-aided translation. *Machine Translation* 23(4):241-263.

Philipp Koehn. 2010. Enabling Monolingual Translators: Post-Editing vs. Options. *Proc. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL.*

Philipp Koehn and Ulrich Germann. 2014. The Impact of Machine Translation Quality on Human Post-editing. *Proc. Workshop on Humans and Computer-assisted Translation.*

Maarit Koponen. 2012. Comparing human perceptions of post-editing effort with post-editing operations. *Proc. 7th Workshop on Statistical Machine Translation*, p.181-190.

Maarit Koponen, Wilker Aziz, Luciana Ramos, and Lucia Specia. 2012. Post-editing time as a measure of cognitive effort. *AMTA 2012 Workshop on Post-Editing Technology and Practice*, p.11-20.

Maarit Koponen. 2013. This translation is not too bad: An analysis of post-editor choices in a machine translation post-editing task. *Proc. MT Summit XIV Workshop on Post-editing Technology and Practice*.

Hans P. Krings. 2001. *Repairing texts: Empirical investigations of machine translation post-editing process*. Kent State University Press.

Philippe Langlais, Sébastien Sauvé, George Foster, Elliott Macklovitch, Guy Lapalme. 2000. Evaluation of TRANSTYPE, a Computer-aided Translation Typing System: A comparison of a theoretical- and a user- oriented evaluation procedures. *Proc. LREC.*

Samuel Laubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. 2013. Assessing post-editing efficiency in a realistic translation environment. *Proc. MT Summit XIV Workshop on Post-editing Technology and Practice*.

Robert C. Moore. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation*, p. 135-144.

Sharon O'Brien. 2011. Towards predicting post-editing productivity. *Machine Translation* 25:197-215.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29(1):19-51.

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

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proc. 7th Conference of the Association for Machine Translation in the Americas*, p. 223-231.

Yan Song, Prescott Klassen, Fei Xia and Chunyu Kit. 2012. Entropy-based Training Data Selection for Domain Adaptation. *Proc. COLING 2012*, p.1191-1200.

Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi and Nello Cristianini. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. *Proc. 13th Annual Conference of the EAMT*, p. 28-35.

Midori Tatsumi and Johann Roturier. 2010. Source Text Characteristics and Technical and Temporal Post-Editing Effort: What is Their Relationship? *Proc. 2nd Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry" (JEC).*

Irina Temnikova. 2010. Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment. *Proc. LREC*.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, Shuo Li, Yiming Wang, Yi Lu. 2014. UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. 2013. Evaluation of machine translation and its evaluation. *Proc. MT Summit IX.*

Lucia Morado Vazquez, Silvia Rodriguez Vazquez, and Pierrette Bouillon. 2013. Comparing forum data post-editing performance using phrase table and machine translation output: a pilot study. *Proc. XIV Machine Translation Summit*.

V. Zhechev. 2012. Machine translation infrastructure and post-editing performance at Autodesk. *Proc. AMTA Workshop on Post-editing Technology and Practice.*