Assessing Post-Editing Efficiency in a Realistic Translation Environment

Samuel Läubli[1] **Mark Fishel**[1] **Gary Massey**[2] **Maureen Ehrensberger-Dow**[2] **Martin Volk**[1]

[1] Institute of Computational Linguistics
University of Zurich
Binzmühlestrasse 14
CH-8050 Zürich

[2] Institute of Translation and Interpreting
Zurich University of Applied Sciences
Theaterstrasse 15c
CH-8401 Winterthur

{laeubli,fishel,volk}@cl.uzh.ch {mssy,ehre}@zhaw.ch

Abstract

In many experimental studies on assessing post-editing efficiency, idiosyncratic user interfaces isolate translators from translation aids that are available to them in their daily work. In contrast, our experimental design allows translators to use a well-known translator workbench for both conventional translation and post-editing. We find that post-editing reduces translation time significantly, although considerably less than reported in isolated experiments, and argue that overall assessments of post-editing efficiency should be based on a realistic translation environment.

1 Introduction

Machine translation has had a considerable impact on the translation industry in recent years, and there are a number of studies that systematically test the efficiency of replacing manual translation "from scratch" with post-editing—the process of manually adapting machine translations. Most of such studies isolate participants from access to additional translation tools since they would affect the timing and other response variables. To ensure precise measurements, translators usually operate via a simple user interface, specially tailored for such studies.

This means, however, that the conditions under which some studies are conducted differ from the final working conditions in which post-editing will be used. The interfaces used are not directly comparable to the translation aids and workbenches currently available to translators in their daily work, which can result in overestimations of time gains through post-editing. We believe that assessments of post-editing efficiency should instead be based on a more *realistic environment* for participating translators.

In this paper, we employ a customary translation workbench to evaluate the effectiveness of post-editing translations of marketing texts from the automobile industry. We hypothesize that providing translators with a domain-specific translation system will increase their productivity even when added on top of other well-known translation aids, such as translation memories and bilingual terminology databases. In line with recent work on inferential statistics in post-editing research by Green et al. (2013), we test whether this productivity increase is statistically significant.

In the following section, we outline our use case and briefly review a number of post-editing studies that have been conducted so far. In Section 3, we detail our experimental design, outlining how we measure translation throughput with and without post-editing (Section 4), as well as the quality of all translations produced in the study (Section 5). In Section 6, we compare the respective results to related studies and contrast them with user perceptions, and then we draw conclusions and outline future work in Section 7.

2 Background

The research reported here is part of a joint project between the University of Zurich and a language service provider (LSP) with a primary focus on translating material from the automobile industry, such as brochures and other marketing texts—a specific domain with its own terminology and typical translations (see Table 1). The aim of the

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. 83–91.

Source (DE)	**Reference** (FR)	**TM-Only** (FR, P4)	**Post-Edit** (FR, P6)	**English gloss**
Streifenbeklebung auf Frontklappe und über den Seitenschwellern	Bande adhésive sur le capot et sur les seuils latéraux	Bandes autocollantes sur le capot et sur les jupes latérales	Bandes adhésives sur le capot et au niveau des ailes	*Adhesive stripes on the bonnet and above the side skirts*

Table 1: A German source segment in translation task D (product features) and its translations into French: A reference translation produced by a specialized translator prior to our experiment and translations by participants in the TM-Only and Post-Edit conditions as well as the English gloss.

project is to build a domain-specific machine translation system for the LSP to use in a post-editing scenario.

There have been a number of studies that assess the efficiency of post-editing (e.g., Guerberof, 2009; Sousa et al., 2011; Green et al., 2013). Most of these set up controlled environments for their experiments and develop specially tailored user interfaces for post-editing tasks (e.g., Aziz et al., 2012). The main reason for this is the priority placed on precise measurements of translation time, pause durations and input device activities.

Some evaluations of post-editing in an industrial context have also been reported. For example, Plitt and Masselot (2010) replace manual translation with post-editing for software localization. Other industry-oriented studies (Volk et al., 2010; Flournoy, 2011) focus more on the challenges of deploying machine translation in the respective sector or company.

Our work strives to combine the key elements of these two approaches: We ensure precise time and activity measurements while preserving a realistic translation environment.

3 Experimental Design

The main idea behind our experiments is to assess the efficiency of post-editing in a realistic translation environment. The participating translators were asked to translate a number of texts in two conditions using Across Personal Edition[1]. The setup for the `TM-Only` condition included access to a translation memory (TM) with 176,957 domain-specific entries. Exact matches of the TM were automatically inserted into the otherwise empty translation template, and fuzzy matches were displayed in a dedicated section of the workbench. In addition, the participants were able to access a small domain-specific terminology database (704 entries) as well as any additional translation aids of their choice, such as printed or online dictionaries.

In the `Post-Edit` condition, machine-translated output was included in addition to the previously described setup, while access to the same translation aids was allowed. However, whereas in the `TM-Only` condition text fields with no exact match in the TM were left empty, they were filled with machine translations (MT) in the `Post-Edit` condition. Machine-translated segments were marked as such, so that the origin of the translation was transparent to the translators. The participants were asked to translate the German source text (`TM-Only`) or revise the French MT output (`Post-Edit`) as needed to produce high-quality French target texts. They were encouraged to work however they wanted and had access to the fuzzy TM matches, the terminology database, and online resources.

Pre-edit translation drafts were produced by a domain-specific statistical machine translation system. It was built using the same translation memory data that was used for the present study, as well as out-of-domain parallel corpora; a more detailed description can be found in Läubli et al. (2013a,b). We implemented a simple RPC-based software link to enable seamless integration of the translation system into the translation workbench.

The German source texts for the translation tasks were provided by the LSP. We selected four typical texts (A–D) that cover specific aspects of our domain in scope (see Table 2). Since all of the texts had been translated by professional translators at the LSP in their normal workflow, we also had access to the corresponding reference translations into French. This allowed us to compare the output of translators experienced in the automobile industry text domain (the LSP staff) with that of those new to it (participants of this study), as well as to assess the effect of post-editing on this comparison.

[1] `http://www.my-across.net/en/translation-workbench.aspx`

Text	**A**	**B**	**C**	**D**
Type	company portrait	letter	presentation slide	product features
Language	full sentences	full sentences	bullet points	bullet points
Segments	7	18	12	13
Words	107	103	50	64
Characters	890	742	489	557
Coverage	poor	good	poor	good
100%	-	3	1	3
80–99%	-	6	-	5
50–79%	1	3	1	2
No match	6	6	10	3

Table 2: Text materials. We used four typical marketing texts from the automobile industry.

The six participants of the present study (P1-P6) were native speakers of the target language (French) and highly competent in the source language (German), between 20 and 40 years of age (mean: 25.5, median: 22.5). All of them were familiar with translation workbench technology and were majoring in German-French translation in a BA program in general translation or in an MA program in specialized translation. Four participants reported regularly translating texts for payment, and two of them had already been employed as professional translators. The participants were compensated for their involvement in the study according to customary rates at their home institute.

Each of the six participants (P1–P6) translated all four texts (A–D) after a familiarization session with the Across translation workbench system. Participant-document assignment was done randomly with three constraints, to guarantee that:

(i) no participant was presented with the same document twice, in any setup;

(ii) each document was translated three times in `TM-Only` and three times in `Post-Edit`;

(iii) each participant translated two documents in each condition.

The time needed for completing the translation tasks was measured by means of screen recordings. This technique, commonly used in translation process research (e.g., Ehrensberger-Dow and Massey, 2013), clearly distinguishes our experimental design from previous studies as the use of screen recordings allows precise time measurements to be made without the need for idiosyncratic user interfaces (such as Aziz et al., 2012) or relying solely on what participants themselves track and report (as, for example, in Plitt and Masselot, 2010).

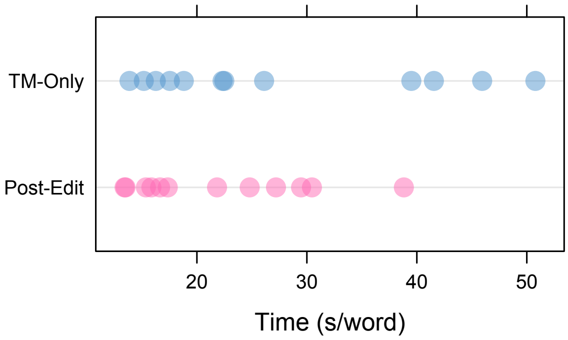

Figure 1: Translation time (seconds per word) by condition.

In the following, we describe the results of efficiency estimation experiments (Section 4), before assessing the quality of the translations produced in the two conditions (Section 5).

4 Translation Throughput

As mentioned above, no document was translated more than once by the same participant. Although this guarantees independent time measurements (translating a document in either condition would have inevitably affected translation timing in the other condition), this also means that we cannot compare those time measurements directly due to several variables such as average translation speed per participant, document length and complexity, etc. Instead, we normalized the measurements.

The standard approach is to normalize the time by the length of the document: i.e., to divide the time per translation task by the number of sentences or words in its source text. However, this only accounts for the length of the document and not for other text characteristics that are more difficult to control in an experimental setting (cf. Table 2). In our data, length-normalized translation times vary considerably by document (see Ta-

Text	Time (s/word)		Change
	TM-Only	Post-Edit	
A	16.0 ± 2.6	16.2 ± 1.0	1.5%
B	18.8 ± 3.1	14.5 ± 1.8	-22.6%
C	45.4 ± 5.7	31.4 ± 7.1	-30.9%
D	30.0 ± 10.2	26.2 ± 3.9	-12.8%

Table 3: Average translation time and standard deviation (seconds per word) by document. Each document was translated by three randomly assigned participants per condition.

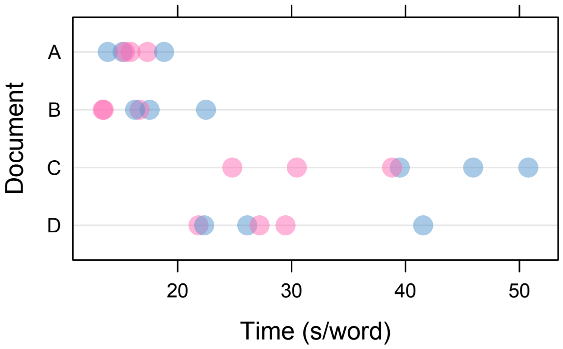

Figure 2: Translation time (seconds per word) by document. Blue and pink data points represent TM-Only and Post-Edit, respectively.

ble 3), indicating that other factors[2] influence the average time needed for translating a word.

Overall, translating a word took participants 27.5 seconds in TM-Only and 22.1 seconds in Post-Edit on average (-19.9%). Standard deviations are high in both conditions (see Figure 1), but clearly higher for TM-Only (± 13.2 seconds) than for Post-Edit (± 8.1 seconds). Translation speed differs greatly between both participants and documents (see Figure 2). The more prose-like texts consisting primarily of full sentences (A, B) were translated much faster than the information-denser texts consisting primarily of bullet points (C, D), regardless of whether the TM coverage was good or poor.

According to the length-normalized measurements, post-editing helped four out of six participants translate faster. However, performing a *per subject* analysis is not appropriate in our setting as each participant translated two texts per condition and individual averages per participant and condition are based on two time measurements only. Results of a *per item* analysis (Table 3) show that three out of four texts were translated faster in Post-Edit; however, the same criticism applies to these results, since time measurements are averaged over only three participants per condition and document.

Looking at our data by participants *or* items is not satisfactory. As we seek to generalize from samples to populations—ideally, all possible translators rather than P1–P6 and all automobile marketing texts rather than texts A–D—we are interested in assessing whether Post-Edit is faster than TM-Only given random variation from both participants and documents. We thus want to test for a genuine difference between our conditions despite extraneous or potentially confounding variables that we cannot fully control in our experiment, such as different translation speed between participants or the frequency of certain word classes in texts.

We thus employed linear mixed effects (LME) models (Baayen et al., 2008) to analyze our data.[3] Green et al. (2013) showed that LME models are preferable to, e.g., analysis of variance (ANOVA) in post-editing experiments because language can be treated as a random effect, thus avoiding the "language-as-fixed-effect fallacy" (Clark, 1973). Accordingly, we only use the translation condition as a fixed effect and both participant and text as random effects. We did not apply a prior normalization of times by text length since length is an implicit feature of the respective random effect. We checked for homogeneity and normality in our data and tested the validity of the mixed effects analyses by comparing the models with fixed effects to null models (comprising only the random effects) through likelihood ratio tests.

The resulting LME model shows a significant main effect for translation condition (MCMC-estimated p-value = 0.0192). The estimated average translation times are 1,957.4 seconds per text for TM-Only and 1,617.7 seconds per text for Post-Edit; i.e., post-editing reduces time by 17.4%.

[2] For example, Green et al. (2013) report a significant correlation between translation time and percentage of nouns in the source text. In our experiment, participants also required more time to translate texts written in nominal style, i.e., bullet points rather than full sentences.

[3] Using the lme4 (Bates et al., 2013) and languageR (Baayen, 2011) packages in R (R Core Team, 2013).

5 Translation Quality

Time savings through post-editing are only relevant if the quality of the produced translations remains consistent. We consider this criterion to be met if the target audience cannot distinguish post-edited from conventionally translated texts, which we tested with translation experts (Section 5.1) and with the participants of our study (Section 5.2).

5.1 Expert Rating

To assess overall translation quality, we asked two independent experts to evaluate all translations in detail. We included the reference translation provided by the LSP, resulting in seven translations per source text—six produced by P1–P6 (three in `TM-Only`, three in `Post-Edit`) and one produced by a professional translator. The experts were not informed about the origins of the translations or the translation conditions. They were provided with the four German source texts (A–D) and the seven French translations of each in unified formatting and random order. Both experts returned the assessments of the 28 translations within a week of receiving them and then had no further involvement in the study. They were compensated according to regulations of the Zurich University of Applied Sciences (ZHAW).

The experts, lecturers in German-French Translation, used the ZHAW's internal evaluation scheme, consisting of five ordinal scales for (i) target language expression, (ii) target language grammar, (iii) target language syntax, (iv) semantic accuracy, and (v) translation strategy. For every translation that the experts rated, they were asked to score each main category on a scale of 1 to 4 points each; the higher the score, the better the performance in that category was considered to be. The five separate scores were aggregated to obtain a value out of 20 for each translation.

The averaged expert ratings reveal clear differences between the `TM-Only` and `Post-Edit` conditions for the translations of the texts containing fully formed sentences (A, B; see Table 4). With an average total score of 15.7, the post-edited translations of text A were rated 16% higher than the `TM-Only` target texts, which scored only 13.5 on average. The post-edited translations of text B scored 15.0, 7% higher than the average total (14.0) for the `TM-Only` condition. The distinctions between the conditions for texts C (`TM-Only`: 13.8, `Post-Edit`: 14.2) and D (`TM-Only`: 16.0, `Post-Edit`: 15.7) are less (2.4% and -2.1% respectively), conceivably because coherence is less of an issue in texts almost exclusively made up of bullet points. It would seem that the `Post-Edit` condition produced full-sentence texts of higher quality.

Text	**Expert Rating**			**Diff.**[a]
	`TM-Only`	Reference	`Post-Edit`	
A	13.5 ± 1.4	14.0	15.7 ± 1.9	16.3%
B	14.0 ± 1.9	13.5	15.0 ± 1.0	7.1%
C	13.8 ± 1.3	15.5	14.2 ± 1.9	2.4%
D	16.0 ± 1.7	15.5	15.7 ± 1.1	-2.1%

[a] Difference between `TM-Only` and `Post-Edit` averages.

Table 4: Expert ratings (points on an ordinal scale; max=20 points). Scores for `TM-Only` and `Post-Edit` are averages over two independent ratings each for three translations per condition and document. Reference scores are averages of two ratings for one professionally produced translation per document.

This appears to be confirmed by the expert scores for the reference translations. Compared with the translations produced in `Post-Edit`, the reference translations received lower average ratings for the full-sentence texts A (company portrait) and B (letter). In two cases, namely texts B and D, the reference translations also scored worse than some of the target texts written in the `TM-Only` condition.

5.2 Pairwise Ranking

In addition, we tested whether the participants (P1–P6) prefer professional translations produced by the LSP staff over those produced in the study. We applied a pairwise ranking procedure[4] in which evaluators compare two translations $\langle t_1, t_2 \rangle$ of a given segment in the source language and choose the better fit, with ties allowed. We used six random German segments from each text (A–D) and had all participants compare the corresponding professional translation to those produced by all

[4] A similar procedure was used at the 2012 Workshop for Machine Translation (Callison-Burch et al., 2012). In contrast to other human evaluation metrics such as fluency and adequacy judgments on ordinal scales, pairwise rankings are usually more comprehensible and better reproducible for non-expert evaluators.

Condition	Wins	Ties	Losses	p-value
`TM-Only`	112	94	154	**0.012**
`Post-Edit`	128	96	136	0.667

Table 5: Pairwise ranking of translations produced in the study against professional reference translations. p-values indicate genuine differences between the number of wins and losses (Sign Test).

other participants, such that each participant evaluated 120 $\langle t_{professional}, t_{participant_i}\rangle$ tuples in total. Participants were not told about the origin of the translations to be compared, i.e., they did not know that they were comparing professional translations to those produced in the study. We presented the translation alternatives in random order and inserted 10 "spam" items per participant—tuples where one translation did not match the original segment—to control for deliberate choices.

Results of the ranking are presented in Table 5. When comparing the LSP's translations to those that have been produced in the `TM-Only` condition, participants preferred the former: The reference translations were preferred in 154 out of 266 non-tie cases (57.9%); the translations by P1–P6 in 112 cases (42.1%). In contrast, reference and participants' translations were rated comparably in the `Post-Edit` condition: The former were preferred in 51.5% of the cases, the latter in 48.5%.

We applied the Sign Test to determine whether the win:loss ratio between `TM-Only` and the professionally produced reference translations (hereafter "Reference") as well as that between `Post-Edit` and Reference is genuine. As presented in Table 5, `TM-Only` is ranked significantly lower than Reference, while the difference between `Post-Edit` and Reference is attributable to chance. In other words, participants could not distinguish their post-edited translations from the professionally produced translations, while they considered the professional translations better than those produced in the `TM-Only` condition.

6 Discussion

Post-editing reduces time significantly even when a fully-featured translation workbench is available. Our results suggest that actual time savings lie within a range of 15–20%, which is, however, considerably lower than numbers reported in other studies. For example, Sousa et al. (2011) found a "speeding-up [of] the translation process by 40%" for film subtitles from English to Portuguese. Plitt and Masselot (2010) report average time savings of 43% in software localisation from English to French, Italian, German, and Spanish.

One reason for this considerable difference may be that we did not recruit professional translators for our study. When compared to the results of Plitt and Masselot, who employed specialist translators with considerable experience in software localisation, the student participants P1–P6 needed a relatively long time to translate in both conditions (see Section 4 and Table 3). However, Sousa et al. obtained time savings of 40% even with volunteers that only "have some experience with translation tasks". In contrast, our participants are pursuing and/or have completed university degrees in professional translation, and most of them regularly translate texts for payment (see Section 3).

In addition to other factors such as text genre and language pairs, the realistic translation environment might explain our high per-word translation time as well as the lower productivity gains. In contrast to many other studies (see Section 2), our translators were not forced to translate texts strictly segment by segment, since they were presented with a complete source text for each task rather than isolated sentences, and translated documents could be revised as a whole before submission. Inspections of screen recordings reveal that participants made extensive use of this possibility, which is common practice among professional translators (Guerberof, 2013). Most importantly, the availability of a domain-specific translation memory and a bilingual terminology database reduced the difference between `TM-Only` and `Post-Edit`, i.e., it increased translation throughput, especially in the former condition, where no machine translations were available.

Finding reasons for *why* translators still work significantly faster in the `Post-Edit` condition was not the focus of our study. However, a preliminary analysis of screen recordings supports Green et al.'s (2013) finding that translators draft less when post-editing. We also noticed that participants often neglected suitable fuzzy translation memory matches in the `TM-Only` condition. When the same or very similar translations were automatically inserted into the target language template in Post-Edit, participants often accepted them with no or only minor changes. It seems that machine translations help translators by providing a clear "starting point", thus eliminating the need for browsing through all available resources (translation memories, websites, or dictionaries) before actually starting to draft.

The evaluation of translation quality (see Section 5) confirms that post-edited translations are at least equivalent to conventionally produced translations. It also highlights the importance of considering the genre, information density, and linguistic structures of source texts when comparing the efficiency of various translation aids. Prose-like texts, such as company profiles and letters, may not be translated faster with MT input, but the final result may be of better quality overall because the translator can focus on editing the text to suit its purpose rather than focusing on translating words and structures. On the other hand, it takes much longer to translate information-dense texts (such as those consisting primarily of bullet points) from scratch, which is why good-quality MT input can help so much (up to 30.9%; see Table 3). With these types of informative texts, editing for cohesion and linguistic style is much less important.

The status of reference translations has been called into question by the findings reported in Section 5.1: Expert markers unaware of the translators' background or training consistently categorized the reference translation as average quality compared with students' translations. However, the comparison is somewhat unfair, since the conditions for the translations were not the same. Professional translators are subject to many constraints, such as time pressure and adherence to clients' style guides, which were not imposed in the present study. From this point of view, a better design for the pairwise ranking might have been to compare the TM-Only and Post-Edit segments not only to the reference translation but also to each other.

A survey that followed the translation tasks revealed that our participants were considerably reserved towards machine translation and post-editing. Five out of six participants considered pre-translations in the Post-Edit condition to be "not useful" (4) or "not at all usefull" (1); only one participant found them "sometimes useful". Overall, five participants preferred to work in the TM-Only condition, while one (P5) preferred Post-Edit, stating

> For the very technical parts of the catalogue [to be translated] I would probably prefer the mode with pre-translations.

P3 indicated that the machine translations were helpful for translating difficult texts in terms of vocabulary,

> [...] but I think [for translating] the two easiest texts [...], the pre-translations would have only confused me.

This stands in sharp contrast to the fact that post-editing resulted in significantly faster and even slightly better translations in our study. However, the discrepancy between translators' perceptions and post-editing performance is a well-known phenomenon in the field (see, e.g., Koponen, 2012). On the other hand, our participants were by no means technology-averse in general: All of them used various computer-aided translation tools in the tasks and deemed both the domain-specific translation memory and the bilingual terminology database as either "sometimes useful" (3/3), "useful" (1/2), or "very useful" (2/1).

7 Conclusion

We have proposed a design for translation efficiency experiments that compares post-editing to computer-aided translation using a fully-featured translation workbench. In contrast to the simplified user interfaces deployed in other studies (e.g., Sousa et al., 2011; Green et al., 2013), this allows participants to use translation memories and terminology databases, long indispensable tools for professional translators. Precise time measurements were obtained by means of screen recordings, which is unobtrusive to participants and eliminates the need for them to track and report times themselves (as in Plitt and Masselot, 2010).

Applying the proposed methodology in a controlled experiment, we have shown that post-editing results in significantly faster translation with consistent quality even when compared to computer-aided translation (as opposed to completely unaided translation). While time savings are most noticeable in dense documents consisting of bullet points, post-editing also facilitated the translation of prose-like texts that require editing for cohesion and linguistic style.

Specialist ratings as well as a pairwise ranking procedure confirm that the quality of post-edited texts is consistent with or, in some cases, even better than conventionally produced translations. Quality improvements through post-editing were mostly found in coherent full-sentence texts.

Overall, our results indicate that gains in translation throughput are around 15–20%, which is considerably lower than numbers reported in studies that isolate participants from commonly used translation tools such as translation memories and bilingual terminology databases. While such isolated studies are clearly important for examining specific aspects of post-editing, our findings strongly suggest that its overall efficiency should be assessed in a realistic environment that takes account of the various aids available to translators in their daily work.

Acknowledgements

We would like to thank Annina Meyer for her substantial help in organizing the quantitative experiments described in this paper. Our work was partly supported by Grant No. 11926.2 PFES-ES from the Swiss Federal Commission for Technology and Innovation.

References

Wilker Aziz, Sheila Castilho, and Lucia Specia. Pet: A tool for post-editing and assessing machine translation. In *Proceedings of LREC*, Istanbul, Turkey, 2012.

R. Harald Baayen. *languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics".*, 2011. URL `http://CRAN.R-project.org/package=languageR`. R package version 1.4.

R. Harald Baayen, Douglas J. Davidson, and Douglas M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412, 2008.

Douglas Bates, Martin Maechler, and Ben Bolker. *lme4: Linear mixed-effects models using S4 classes*, 2013. URL `http://CRAN.R-project.org/package=lme4`. R package version 0.999999-2.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of WMT*, pages 10–51, Montréal, Canada, June 2012.

Herbert H. Clark. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4):335 – 359, 1973.

Maureen Ehrensberger-Dow and Gary Massey. Indicators of translation competence: Translators' self-concepts and the translation of titles. *Journal of Writing Research*, 5:103–131, 2013.

Raymond Flournoy. MT use within the enterprise: Encouraging adoption via a unified MT API. In *Proceedings of MT Summit XIII*, pages 234–238, Xiamen, China, 2011.

Spence Green, Jeffrey Heer, and Christopher D. Manning. The efficacy of human post-editing for language translation. In *Proceedings of CHI*, Paris, France, 2013.

Ana Guerberof. Productivity and quality in the post-editing of outputs from translation memories and machine translation. *International Journal of Localisation*, 7(1):11–21, 2009.

Ana Guerberof. What do professional translators think about post-editing? *Journal of Specialised Translation*, 19:75–95, 2013.

Maarit Koponen. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of WMT*, pages 181–190, Montréal, Canada, 2012.

Samuel Läubli, Mark Fishel, Martin Volk, and Manuela Weibel. Combining domain-specific translation memories with general-domain parallel corpora in statistical machine translation systems. In *Proceedings of NODALIDA*, pages 331–341, Oslo, Norway, 2013a.

Samuel Läubli, Mark Fishel, Manuela Weibel, and Martin Volk. Statistical machine translation for automobile marketing texts. In *Proceedings of MT Summit XIV*, Nice, France, 2013b.

Mirko Plitt and François Masselot. A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague*

Bulletin of Mathematical Linguistics, 93:7–16, 2010.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

Sheila C.M. de Sousa, Wilker Aziz, and Lucia Specia. Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In *Proceedings of RANLP*, pages 97–103, Hissar, Bulgaria, 2011.

Martin Volk, Rico Sennrich, Christian Hardmeier, and Frida Tidström. Machine translation of TV subtitles for large scale production. In *Proceedings of EM+/CNGL*, pages 53–62, Denver, USA, 2010.