PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 93 JANUARY 2010 7–16

# A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context

Mirko Plitt, François Masselot

Autodesk Development Sàrl, Neuchâtel, Switzerland

**Abstract**

We evaluated the productivity increase of statistical MT post-editing as compared to traditional translation in a two-day test involving twelve participants translating from English to French, Italian, German, and Spanish. The test setup followed an empirical methodology. A random subset of the entire new content produced in our company during a given year was translated with statistical MT engines trained on data from the previous year. The translation environment recorded translation and post-editing times for each sentence. The results show a productivity increase for each participant, with significant variance across inviduals.

## 1. Introduction

The machine translation productivity test described in this article was conducted in the context of the deployment of machine translation at Autodesk, a software company whose products are translated ("localised") from English into up to twenty languages. We held this test to manage expectations as to the financial savings our company would be able to achieve thanks to machine translation.

Publicly available data on post-editing productivity of statistical machine translation in localisation is scarce (O´Brien, 2005; Takako et al., 2007; Schmidtke, 2008; De Sutter et al., 2008; Flournoy and Duran, 2009). Furthermore, most of the data that is available has not been acquired under controlled conditions (Krings, 2001).

Specific limitations of other post-editing productivity tests that prevented us from using their results included:

- Unclear test objectives leading e.g. to non-representative training corpora.
- Untypical translator profiles.


Corresponding author: mirko.plitt@autodesk.com
Cite as: Mirko Plitt, François Masselot. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. The Prague Bulletin of Mathematical Linguistics No. 93, 2010, pp. 7–16. ISBN 978-80-904175-4-0. doi: 10.2478/v10108-010-0010-x.

- Artificial test sets (e.g. because of a close relation with the corpus[1], or because text deemed unsuitable for MT was removed);
- Absence of traditional translation benchmarks (e.g. assuming a daily throughput of 2500 words, a common rule of thumb in the localisation industry).
- Unreliable time measurement (e.g. based on times reported by individual participants), if any.
- Commercial bias.

The machine translation system we selected for our productivity test was the open-source Moses system (Koehn et al., 2007), trained solely on our own data without any factored representation.

We chose Moses for the following main reasons: (i) the language-independent nature of statistical machine translation makes it easily expandable across several languages at once; (ii) as a typical translation service buyer we possess considerable amounts of high-quality legacy translations; (iii) it would have been difficult to reach return on investment with a commercial machine translation system.

## 2. Test Setup

The principal aim of our productivity test was to measure the productivity increase we could expect in production at Autodesk. The actual productivity numbers presented in this article may therefore be of limited use for other users of machine translation. Elements of the experimental approach we took to obtain these numbers, however, can be applied beyond our specific case. The following aspects of our approach merit particular attention:

### 2.1. Test Set Selection

We simulated a Moses production deployment in the most recent round of translation.[2] We therefore trained engines on all our translation data up to the end of 2008. The test set was a randomly selected subset of all the new[3] content submitted for translation in *2009*.

The random selection ensured that the test set was representative in every sense, including any phenomenon that may or may not influence MT quality and post-editing productivity. We split the test set into "jobs", grouped by product (to preserve some context), and sentences were kept in their original order (if often separated by gaps).

[1] We believe that the practice of "cutting" a test set from a corpus presents the risk of introducing a bias in the relation between the two.

[2] The majority of Autodesk products are released once per year; translation activity therefore follows yearly cycles.

[3] *New* means, in this instance, sentences yielding translation memory matches below 75%. In a typical localisation scenario, the use of translation memory technology leaves little room for the deployment of MT above this threshold (Bruckner and Plitt, 2001; Carl and Hansen, 1999).

### 2.2. Post-Editing Environment

To measure translation time as precisely as possible, without relying solely on what test participants would track and report back to us, we developed our own "workbench", a post-editing environment largely inspired by the caitra environment (Koehn and Haddow, 2009). The workbench was designed to capture keyboard and pause times for each sentence, and was implemented in Ruby on Rails, a web application framework that readily offered most of the functionality required.

The workbench interface (see Figure 1) presented the source and target sentences one beneath the other. For the post-editing tasks, the target sentence field was prepopulated with the MT proposal, to prevent test participants from translating from scratch. The workbench recorded the edit time, the number of edit sessions and the number of key strokes for each sentence.

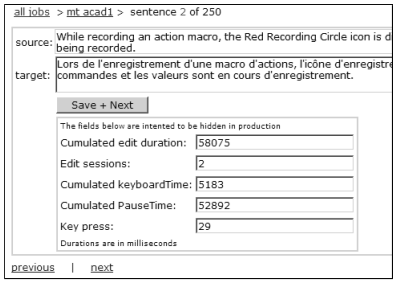


*Figure 1. Workbench screenshot (time recording fields were hidden from translators)*

### 2.3. Test Participants

We chose three of our usual localisation vendors for the test. Each vendor assigned one translator per language. We did not intervene in the translator selection as such, and did not request candidates to present particular profiles in terms of translation speed or quality, or post-editing experience. We did not provide our test participants with any training but gave simple post-editing instructions.

### 2.4. Translation Productivity Benchmark

The productivity test was divided in two phases; the first phase consisted of traditional translation without support from MT—to obtain a reference value for each individual test participant—and only the second phase was dedicated to post-editing. We assigned the jobs in such a way that each translator was to do at least one job in

each of the three product domains, both in post-editing and translation. We also made sure that participants would not translate and post-edit the same job.

### 2.5. Quality Assessment

Our expectation was that the quality of the post-edited translation would be equivalent to traditional translation, quality being defined here according to the standard criteria applied at Autodesk.

To verify that this expectation was met, we provided the Autodesk translation QA team with samples of translated and post-edited text, again randomly selected, and of reasonable size. The QA team was aware of the overall context of the productivity test but did not know which text was the result of post-editing and which was a traditional translation.

### 2.6. Test Execution

The test was scheduled to last two days. The source language was English, and the target languages were French, Italian, German, and Spanish. Given that we had opted for three translators per language, there were a total of twelve test participants.[4] The scope of the test was defined by what we considered the minimum of meaningful data at a reasonable cost compared to the anticipated savings potential in production.

We prepared 96 jobs, of which 75 ended up being processed, some entirely, some only partially. The cross-product of jobs, languages, and translation types corresponds to 144,648 source words processed.

A small number of sentences, 1.6%, had a duration above five minutes and up to three hours, cumulating to a total 22% of the the time recorded, without there being any explanation such as the complexity of the source text. These sentences were removed from the result set.

## 3. Test Results

### 3.1. Throughput

Figure 2 summarises the test results in terms of throughput. It is most interesting to look at the throughput delta between translation and post-editing for a given translator. Absolute throughputs range from 400 to 1800 words per hour.

Variance across translators was high. MT allowed all translators to work faster, though in varying proportions: from 20% to 131%[5]. MT allowed translators to improve their throughput on average by 74%; in other words, MT saved 43% of the translation time.[6]

[4] One translator chose not to correct tag positions in MT proposals. This translator's work was discarded.

[5] where a 100% productivity gain corresponds to doubling the throughput.

[6] $1 - \frac{1}{1+0.74} = 0.43$

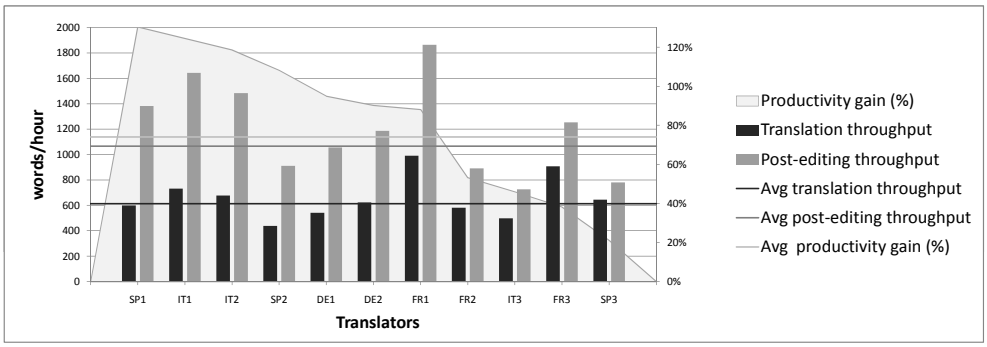


*Figure 2. Individual productivity in words per hour (sorted by descending productivity gain)*

Figure 3 illustrates that in our test, the benefits from MT were greater for slower than for faster translators. Fast translators presumably have a smaller margin of progression because they have already optimised their way of working.

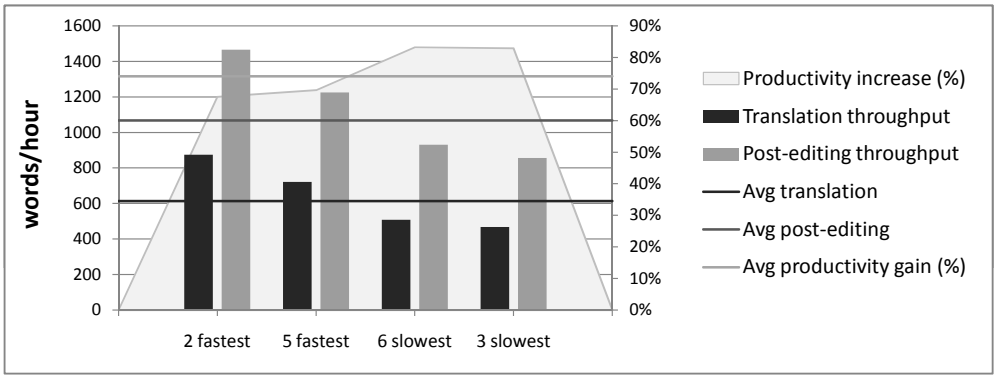


*Figure 3. Fast and slow translators*

### 3.2. Edit Distance and Post-Editing Effort

We calculated edit distances to measure the post-editing effort. We used four different scoring methods: Non-Edited, (sentence-level) BLEU (Papineni et al., 2002), Word Error Rate (WER) (Hunt, 1989; McCowan et al., 2005) and Position-independent

Error Rate (PER) (Tillmann and Ney, 2003). Non-Edited represents the ratio of sentences that were left unchanged. We found that these four indicators, despite their different computing methods, correlate relatively well.

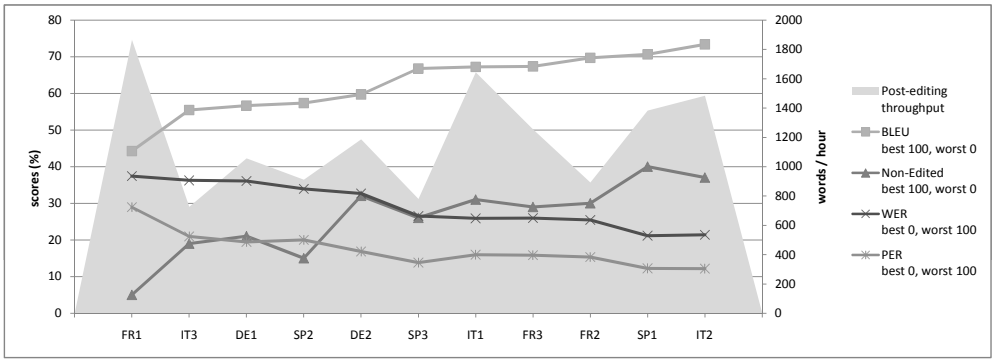


*Figure 4. Post-editing throughput and edit distance (sorted by ascending BLEU score)*

Figure 4 shows a comparison between post-editing throughput and edit distance. One could intuitively expect that fast translators make fewer changes than slow translators. In our test, however, the post-editor who made the highest number of changes was also the fastest. The graphs indicate no clear correlation between edit distance and throughput.

### 3.3. Sentence Length

We also examined the relation between the time spent on sentences and the number of words they contained. Figure 5 shows linear regression for segments up to 35 words.[7]

Figure 6 shows the throughput in *words per hour*, in relation to sentence length. An optimum throughput appears to be reached for sentences of around 25 words. Optima for translation and post-editing are relatively close: around 25 words for translation and 22 words for post-editing. The shapes of the polynomial regression curves indicate that the negative impact of very long sentences on throughput is greater for post-editing.

The productivity gain from post-editing corresponds to the vertical distance between the lines; the optimum is situated around 22 words per sentence. 20–25 word sentences are probably more likely to be semantically self-contained than shorter sentences, thus requiring fewer context checks. The minimal time spent on the translation

[7]Sentences with more than 35 words were infrequent in our test set.

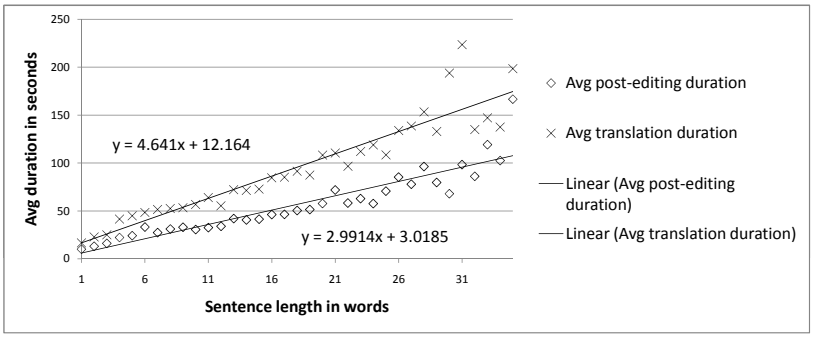


*Figure 5. Average duration and sentence length*

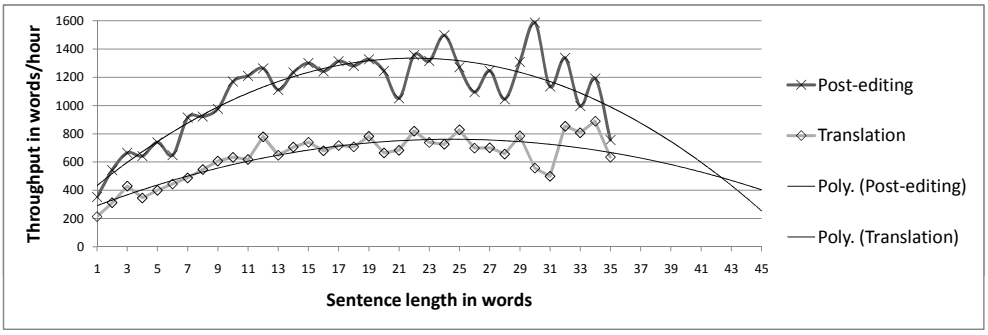


*Figure 6. Words per hour and sentence length*

of any sentence, including the navigation within the text, plays also a proportionally bigger role for shorter sentences.

### 3.4. Influence of Language and Product Domain

Average throughput of translators by language essentially reflects individual differences. Our data does not suggest that MT is more suited for one of the four test languages than for another. We were surprised that the productivity increase of German translators was in line with their French, Italian, and Spanish colleagues, despite the lower quality of the German output that we perceive ourselves.

There was no indication either that the content taken from one product was more suitable, or less, for post-editing than content from other products.

### 3.5. Keyboard Time versus Pause Time

We only recorded two types of editing time: keyboard time and pause time[8]. Pause time can be assumed to include activities such as reading, thinking, and consulting of references.

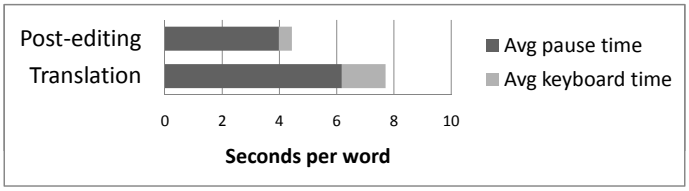


*Figure 7. Keyboard and pause time per word*

Figure 7 shows that keyboard time represents 19% of the edit time for translation and only 10% for post-editing. MT reduces keyboard time by 70% and pause time by 31%. It seems logical that a good MT proposal saves typing time, but it also saves a third of the "thinking" time.

Keyboard and pause time variations were consistent across products, languages and individuals.

### 3.6. Work Regularity

Figure 8 plots, for each job, the standard deviation of the seconds-per-word data series recorded for each sentence. The data suggests that MT evens out the work pace of translators. Our interpretation of this result is that the positive impact of the presence of MT proposals is not only limited to a subset of content or to specific types of sentences.

### 3.7. Quality Assessment

The Autodesk linguistic quality assurance team reviewed part of the jobs of ten of the twelve test participants, evenly split between translation and post-editing jobs for each language. The team rated all the jobs reviewed as either average or good, so all would have been published as is.

The proportion of sentences for which our QA team flagged corrections is grouped in Figure 9. To our surprise, translation jobs contained a higher number of mistakes than post-editing jobs.

[8]keyboard time = sum of time intervals separating two key strokes *inferior* to one second; pause time = sum of time intervals separating two key strokes *superior* to one second

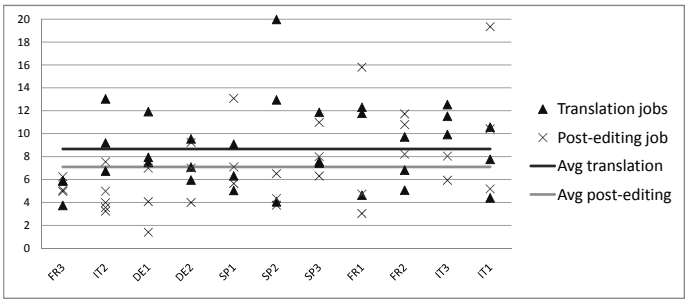


*Figure 8. Standard deviation of seconds per word for each job*

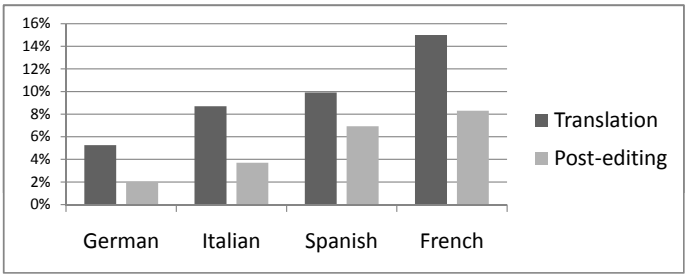


*Figure 9. Percentage of sentences with translation errors*

### 3.8. Translator Feedback

The test participants sent us ample feedback on their experience. On the whole, their comments matched our observations and showed that the test had worked well from their perspective too. However, some of the attempts to interpret their experience were in contradiction with our observations, such as an alleged loss of productivity on longer sentences. There also was contradictory feedback from different participants related to the correctness of product terminology.

## 4. Conclusion

Our test showed that the post-editing of statistical machine translation, when trained and used on Autodesk data, allows translators to substantially increase their productivity. Autodesk has since deployed Moses in production. The empirical methodology followed in the test setup and described in this article can be applied to other real-world evaluations of post-editing productivity.

## Bibliography

Bruckner, Christine and Mirko Plitt. Evaluating the operational benefit of using machine translation output as translation memory input. In *MT Summit VIII, MT evaluation: who did what to whom (Fourth ISLE workshop)*, pages 61–65, Santiago de Compostela, Spain, 2001.

Carl, Michael and Silvia Hansen. Linking translation memories with example-based machine translation. In *Machine Tranlation Summit VII*, pages 617–624, Singapore, Singapore, 1999.

De Sutter, Nathalie, Marie-Laure Poëte, and Joeri Van de Walle. Machine translation productivity evaluation report. Unpublished report on the evaluation of two commercial MT systems conducted for Autodesk, April 2008.

Flournoy, Raymond and Christine Duran. Machine translation and document localization at adobe: From pilot to production. In *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, Ottawa, Ontario, Canada, August 2009.

Hunt, Melvyn J. Figures of merit for assessing connected-word recognisers. In *SIOA-1989*, volume 2, pages 127–131, 1989.

Koehn, Philipp and Barry Haddow. Interactive Assistance to Human Translators using Statistical Machine Translation Methods. In *MT Summit XII*, 2009.

Koehn, Philipp, 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. Moses: Open source toolkit for statistical machine translation. In *ACL Companion Volume. Proc. of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Krings, Hans Peter. *Repairing texts: empirical investigations of machine translation post-editing processes*. Kent State University Press, Kent, Ohio, USA, 2001.

McCowan, Iain, Darren Moore, John Dines, Daniel Gatica-Perez, Mike Flynn, Pierre Wellner, and Hervé Bourlard. On the use of information retrieval measures for speech recognition evaluation. Technical report, Idiap Research Institute, Martigny, Switzerland, March 2005.

O´Brien, Sharon. Methodologies for measuring the correlations between post-editing effort and machine translatability. *Machine Translation*, 19(1):37–58, March 2005.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

Schmidtke, Dag. Microsoft office localization: use of language and translation technology. 2008. URL `http://www.tm-europe.org/files/resources/TM-Europe2008-Dag-Schmidtke-Microsoft.pdf`.

Takako, Aikawa, Lee Schwartz, Ronit King, Mo Corston-Oliver, and Carmen Lozano. Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In *Proceedings of the MT Summit XI*, Copenhange, Denmark, October 2007.

Tillmann, Christoph and Hermann Ney. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Comput. Linguist.*, 29(1):97–133, 2003.