Perception vs Reality: Measuring Machine Translation Post-Editing Productivity

Federico Gaspari fgaspari@computing.dcu.ie
CNGL, School of Computing, Dublin City University, Dublin 9, Ireland
Antonio Toral atoral@computing.dcu.ie
CNGL, School of Computing, Dublin City University, Dublin 9, Ireland
Sudip Kumar Naskar * sudip.naskar@cse.jdvu.ac.in
Jadavpur University, Kolkata, India
Declan Groves * degroves@microsoft.com
Microsoft, Dublin, Ireland
Andy Way away@computing.dcu.ie
CNGL, School of Computing, Dublin City University, Dublin 9, Ireland

Abstract

This paper presents a study of user-perceived vs real machine translation (MT) post-editing effort and productivity gains, focusing on two bidirectional language pairs: English—German and English—Dutch. Twenty experienced media professionals post-edited statistical MT output and also manually translated comparative texts within a production environment. The paper compares the actual post-editing time against the users' perception of the effort and time required to post-edit the MT output to achieve publishable quality, thus measuring real (vs perceived) productivity gains. Although for all the language pairs users perceived MT post-editing to be slower, in fact it proved to be a faster option than manual translation for two translation directions out of four, i.e. for Dutch→English, and (marginally) for English→German. For further objective scrutiny, the paper also checks the correlation of three state-of-the-art automatic MT evaluation metrics (BLEU, METEOR and TER) with the actual post-editing time.

1 Introduction

Machine translation (MT) has developed considerably in the last few years, to the point that it has started to be implemented in industrial translation production scenarios (DePalma, 2011). The industry is embracing MT for certain use-cases, mainly because post-editing (PE) MT output (as opposed to translating from scratch) can lead to productivity gains, particularly within well-defined technical domains (cf. Plitt and Masselot, 2010). However, on the whole, sceptical attitudes remain towards the real benefits of implementing workflows involving MT followed by PE in an effort to speed up the translation process. The main motivation behind this study lies in the need to investigate on the basis of solid evidence the attitudes towards MT PE viz. the actual benefits it yields to obtain high-quality translations. More specifically, the paper examines the extent to which perception matches reality when comparing full MT PE with manual translation from scratch by looking at subjective evaluations (i.e. user perception of time investment, effort and preferred working method) against objective measurements (i.e. the actual time gains).

* Work done while at CNGL, School of Computing, Dublin City University, Ireland.

The rest of the paper is structured as follows. Section 2 examines related work, showing the growing attention that PE has received in the last few years, especially with regard to studies focusing on PE effort and productivity gains; Section 3 outlines the experimental design and set-up of our study, describing the methodology and materials; Section 4 presents the results obtained from the questionnaire that was given to the participants in the study, and Section 5 analyzes the key experimental results concerning real vs perceived productivity gains with PE, also in relation to three state-of-the-art automatic MT evaluation metrics. Finally, Section 6 draws some conclusions, briefly summarizing the key findings of the study.

2 Related Work

Plitt and Masselot (2010) report a productivity test conducted on MT followed by PE as compared to traditional human translation in an industrial environment. They observed that MT helped translators to substantially improve their productivity: MT followed by PE improved throughput on average by 74%, which in effect reduced translation time by 43%. Zhechev (2012) carries out a PE productivity test[1] using a CAT-based PE environment at Autodesk for nine language pairs, and he also found that MT followed by PE results in substantial productivity gains over translation from scratch (ranging from 37% to 92%, depending on the language pair).

Läubli et al. (2013) report experiments carried out in a realistic translation environment and conclude that PE led to significant time gains, even when a fully functional translators' workbench is available. Tatsumi and Roturier (2010) studied the correlation between source-text characteristics and their effects on technical and temporal PE effort on a small English–Japanese dataset. They observed strong correlation between Systran's complexity and ambiguity scores and technical PE effort, and moderate correlation between the IQ score provided by Acrolinx (a widely used authoring software product) and temporal PE effort. Poulis and Kolovratnik (2012) conduct a large-scale evaluation of MT PE aimed at estimating the business benefits of using MT for the European Parliament on 5 European language pairs. They found that on an average 21.4% of the translated segments were rated excellent by the human evaluators (i.e. requiring no PE), while 25.6% of the translations were deemed good (i.e. requiring only minor PE effort). In addition, 20.8% and 32.2% of the translations were found to be average (i.e. requiring major PE) and poor (i.e. of no use), respectively.

Koponen et al. (2012) suggested using PE time as a measure of assessing the cognitive effort involved in PE. They tried to identify different types of MT errors and correlate them with the different levels of difficulty involved in fixing them, where difficulty is measured in terms of PE time. Koponen (2012) studied the relationship between cognitive and technical aspects of PE effort by comparing human scores of perceived effort necessary with the actual edits made by post-editors for cases in which the edit distance and manual scores reflecting perceived effort diverged. The results of an error analysis performed on such data are discussed in terms of the clues that they might provide about edits requiring greater or less cognitive effort compared to the technical effort involved.

Guerberof (2009) studied the effectiveness of using MT output as opposed to translation memory fuzzy matches for the purpose of post-editing in an English→Spanish translation task. She used Language Weaver's statistical MT engine and trained it on the same TM, performing both quantitative and qualitative analyses. The main result was that the productivity of the translators as well as the quality of the translation improved when post-editing MT output, compared to when processing fuzzy matches from the translation memory database.

[1] http://langtech.autodesk.com/productivity.html.

Koehn and Haddow (2009) describe Caitra, a tool that makes suggestions for sentence completion, shows word and phrase translation options, and supports PE of MT output. They report a user study carried out with the tool involving 7 translators for the English–French language pair. Among the different types of assistance offered by Caitra, users prefer the prediction of sentence completion and the options from the translation table over the other types of assistance available for post-editing MT output. To the authors' surprise, PE received the lowest scores among all the options, both in terms of enjoyment and subjective usefulness, although PE was as productive as the other types of assistance.

In an effort to extend the initial insights presented in particular by Koehn and Haddow (2009) and Koponen (2012), this paper investigates perceived vs real productivity gains brought about by post-editing MT output compared against manual translation from scratch in the relatively open – and thus particularly challenging – news-oriented domain.

3 Set-up of the Study

3.1 Methodology and Materials

Output from the CoSyne statistical MT systems (Martzoukos and Monz, 2010) was used in this experiment, and a facility was in place to track the time required by the users to post-edit MT output and to perform manual translations from scratch on texts of similar length and complexity. The texts chosen for the study were extracted from "Today in History/Kalenderblatt" and "Beeld en Geluidwiki", the public wiki sites of the two media organizations that acted as end-user partners in the CoSyne project, namely Deutsche Welle (DW) and the Netherlands Institute for Sound and Vision (NISV).[2] These two bilingual wiki sites cover news, accounts of historical events, biographies of personalities from TV and cinema and descriptions of films and TV series. The texts considered for the experiment were representative of typical source texts used during translation production, as DW and NISV were investigating ways of incorporating MT into their workflows to translate entries of these public wiki sites.

DW chose 10 texts to be translated from German into English with 399 sentences overall, and 10 texts in the opposite direction with 390 sentences. On the other hand, 15 wiki texts were chosen for Dutch to English (394 sentences in total) and 21 for English to Dutch (346 sentences) by NISV. Staff from each organization worked on the respective texts for their own language pair. The quantity of data for each translation direction was roughly similar, with approximately 6,000 words in each of the four source languages. Each sentence was translated manually from scratch and fully post-edited after MT processing by two different participants for the same translation direction. To maximise the amount of data available, each user worked on different texts, taking on the role of experimental subject (using MT followed by PE) and control group member (translating manually from scratch), in turn.

3.2 The Questionnaire

Part of the study was conducted through a preliminary MT evaluation questionnaire given to all 20 participants,[3] which was subsequently supplemented by the collection of experimental PE data. A few initial items in the questionnaire covered basic personal information concerning the age and gender of the respondents, their role in the organization and their professional experience, as well as their previous use of MT. The remaining parts focused specifically on

[2] The URLs are www.todayinhistory.de and www.beeldengeluidwiki.nl, respectively.

[3] The full questionnaire is available at www.computing.dcu.ie/~atoral/resources/questionnaire_post-editing_perception.pdf (only the answers directly relevant to the study are discussed in this paper).

the judgements of the users on the quality of the CoSyne MT output and on their perception of PE compared against manual translation from scratch for translating wiki texts.

At the beginning of the evaluation sessions, the users were informed that for the purposes of this experiment, "post-editing" meant checking the raw output provided by the MT system against the source-language input, revising and improving it as required to obtain a final target text of publishable quality. The purpose of this was to add the final revised translation to the public wiki of their respective media organization; hence, the scenario was that of full PE, aiming for optimal quality of the final revised text (Allen, 2003: 306). In addition, it should be noted that while all participants in the experiment had experience in manual translation, none of them had been specifically trained to carry out PE in a realistic professional task. This is quite different from previous studies such as Snover et al. (2006: 227), where monolingual annotators "were coached on how to minimize the edit rate". To sum up, our study focused on a scenario in which (i) the translators were not trained specifically on PE, and (ii) the objective was publishable quality, as a means of investigating the role of full PE in industrial settings, especially in terms of the perceived vs actual productivity gains.

4 Questionnaire Results

4.1 Profiles of the Participants

At the time of completing the questionnaire, the youngest DW staff member was 38 years of age, and the oldest was 59. Overall, the average age of DW staff who conducted the experiment for the English—German language pair was just over 43 years. In contrast, the NISV employees were aged between 26 and 35, and their average age was just above 31 years. In terms of gender, 7 of the 10 DW respondents were male, and the remaining 3 female. The NISV staff were evenly split between 5 men and 5 women. In total, therefore, the sample of respondents for the two language pairs consisted of 12 men and 8 women. These individuals held a variety of roles within their media organizations (e.g. journalists, editors, etc.), and all of them contributed in various capacities to the creation, development and management of multilingual content on the public wiki sites, from which the texts for this study were extracted. As part of their work, some of these subjects frequently translated content similar to that involved in the experiments conducted for this study in the same language pairs.

In terms of the experience in their organization, DW respondents had been with their employer from a minimum of 1 year (which was the case for a freelance collaborator) to a maximum of 15 years (a senior project manager), with the average being slightly more than 5 years. In contrast, the time spent by NISV staff with their current employer ranged from a minimum of 4 months (in the case of a recently hired professional) to a maximum of 5 years (the chief wiki editor), with the average being just under 3 years of continuous employment.

4.2 Evaluation for EN—DE at DW

This section concerns the questionnaire answers provided by DW staff for the English—German language pair, while Section 4.3 focuses on English—Dutch translation, evaluated by the NISV employees. It should be kept in mind in this respect that in the remainder of this analysis the evaluations for each of the four translation directions under study were formulated by 5 people, supplemented by a comparable control group of the same size.

After performing full PE on the CoSyne MT output to bring it to publishable quality, the users were asked a number of questions focusing on their subjective perception of PE. In particular, the respondents were asked which working method in their opinion involved more effort, i.e. PE of MT output or manual translation from scratch; the lower the score (on a 5-

point Likert scale), the more negative the opinion held by the respondents on the cost-effectiveness of PE compared to manual translation. As shown in Table 1, for the German—English language pair the responses tended to cluster in the lower part of the spectrum, with average scores of 2.0 for the translation direction into English, and 1.75 for translations into German.[4]

Translation direction	*MT with post-editing*				*Manual translation*	*Don't know*	***Avg. (1-5)***
DE→EN	2	2		1			***2.0***
EN→DE	2	1	1			1	***1.75***

Table 1. Effort perception: MT with post-editing vs manual translation for EN—DE.

Next, the users were asked to comment on which working method they thought was faster (thus relying on their own perception) for the two translation directions, i.e. post-editing MT output or manual translation from scratch. The answers are summarized in Table 2, and it is clear that for the English—German language pair there was a strong perception that manual translation from scratch was faster than PE.

Translation direction	*MT with post-editing*				*Manual translation*	*Don't know*	***Avg. (1-5)***
DE→EN		1		2	2		***4.0***
EN→DE			1	2	2		***4.2***

Table 2. Speed perception: MT with post-editing vs manual translation for EN—DE.

Finally, the questionnaire asked which working method the users preferred between post-editing MT and translating from scratch. Table 3 presents the answers given by DW staff.

Translation direction	*MT with post-editing*		*Manual translation*	***Avg. (1-3)***
DE→EN	2	1	2	***2.0***
EN→DE	1		4	***2.6***

Table 3. Overall preference: MT with post-editing vs manual translation for EN—DE.

In the case of German→English translations, there is a neutral situation, with the average score being 2 out of a 3-point scale: 1 respondent had no preference (middle column), while the other two pairs of participants expressed opposite opinions. This might suggest that personal predisposition and possibly expectations related to MT quality could be playing a role in this area. Interestingly, in the opposite direction (English→German) there is a marked preference for manual translation from scratch, with an average score of 2.6 out of 3. However, this opinion was not unanimous, because 1 out of the 5 interviewees stated that they preferred post-editing MT output rather than translating manually from scratch, again pointing to the role of subjective variability in this area.

An important point must be added in this respect, which also applies to the EN—NL language pair, analyzed in Section 4.3, namely that the best human translators are not necessarily the best post-editors. This is particularly relevant here, given that our experimental sub-

[4] In Tables 1-6, for each language pair, the figures in the right-most "Avg." column indicate the average scores (out of 5 or out of 3, as shown). The integer numbers in the remaining columns show how many of the 5 evaluators for that language pair provided the relevant response between the two extremes (the columns in between corresponding to intermediate values along the Likert scales, with the middle one representing neutral "neither one, nor the other" answers). Empty cells mean that no respondents provided that answer, while the "Don't know" column records the number of respondents who did not have a clear answer on the specific point.

jects had not received any specific training in PE, and were naive to the task, while they had varying levels of experience in traditional translation; similarly, there is no reason to suggest that a negative opinion, or a preconceived dislike, of PE leads to a poor PE performance, e.g. one that is slower than human translation from scratch. Due to lack of space, in this paper we do not investigate the relationship between these dimensions concerning the perception and the reality of PE productivity gains, but these are issues that deserve further study.

4.3 Evaluation for EN—NL at NISV

This section concerns the questionnaire results for English—Dutch, and follows the same structure of Section 4.2 for ease of comparison with the results analyzed for the English—German language pair. With regard to the perceived effort that the NISV participants associated with using MT output followed by PE to translate wiki entries, as opposed to manual translation from scratch, Table 4 shows the results for translations from Dutch, where the overall score is slightly in favour of MT with PE (2.75 out of 5 points, with one "don't know" answer); however, the opposite is true for translations into Dutch, for which the users clearly attribute much more effort to MT followed by PE.

Translation direction	*MT with post-editing*				*Manual translation*	*Don't know*	***Avg. (1-5)***
NL→EN		2	1	1		1	***2.75***
EN→NL	4	1					***1.2***

Table 4. Effort perception: MT with post-editing vs manual translation for EN—NL.

The NISV employees were also asked which of the two working methods they perceived to be faster, and the answers to this question are summarized in Table 5.

Translation direction	*MT with post-editing*				*Manual translation*	*Don't know*	***Avg. (1-5)***
NL→EN	1	2		1		1	***2.25***
EN→NL				1	3	1	***4.75***

Table 5. Speed perception: MT with post-editing vs manual translation for EN—NL.

There is a slight preference for using MT followed by PE in the NL→EN translation direction: the average score in that case is 2.25 out of a 5-point scale. However, the opposite is true in the other translation direction, with a total score of 4.75 out of 5.0, and again 1 respondent who opted for "don't know". This means that in general for translations into Dutch, manual translation from scratch was thought to be much less time-consuming than post-editing MT output.

Finally, NISV users were asked about their overall preference between post-editing MT output on the one hand and manual translation from scratch on the other to translate wiki texts between Dutch and English, and Table 6 shows the answers in this respect.

Translation direction	*MT with post-editing*		*Manual translation*	***Avg. (1-3)***
NL→EN		1	4	***2.8***
EN→NL		1	4	***2.8***

Table 6. Overall preference: MT with post-editing vs manual translation for EN—NL.

For both translation directions there was the same overall score, clearly in favour of manual translation from scratch, i.e. 2.8 on a 3-point scale. The strong tendency to indicate manual translation as the preferred working method for English—Dutch seems to be less influenced by personal inclinations, with 1 respondent having no preference in both cases (recorded in the middle column), but all the other members of the sample favouring manual translation.

5 Analysis of Experimental Results

Following on from the subjective evaluation presented in Section 4, this section focuses on the objective experimental results. In particular, in Section 5.1 we zoom in on the actual user performance, comparing real versus perceived PE time and productivity gains for all language pairs. To add a further objective dimension to this part of the study, in Section 5.2 we calculate the correlation between actual PE time and the scores of three state-of-the-art automatic evaluation metrics for the MT output of all four translation directions.

5.1 Time Tracking: Real vs Perceived Productivity Gains with PE

During the evaluation sessions, timestamps were recorded for the manual translation from scratch of the documents as well as for MT PE; we can therefore compare the time taken when translating from scratch with the time spent post-editing MT output, normalizing the measure to the average per single word on the source side. These measures allow us to objectively quantify any productivity gains that are achieved with PE for each translation direction; this can, in turn, be compared with the users' perceptions of time gains.

Following Plitt and Masselot (2010) and Zhechev (2014: 9), we filtered out from our analysis the data related to the texts for which the translation took substantially more time than the others, on the basis of the average processing time per source-language word. We decided to discard the texts for which the processing time (either translation from scratch or PE) took more than 7 seconds per word, which were considered outliers for our purposes. This corresponds to only 2 texts, both for the German→English translation direction, accounting for 3.6% of the overall data sets, but which consumed a notably higher proportion (9.1%) of the overall translation time; we therefore did not want these outliers to unduly skew the results.

Figure 1 shows the average time taken to translate the documents for each translation direction (measured in seconds per word, calculated on the source language), both when translating from scratch (columns HT) and when post-editing MT output (columns PEMT).

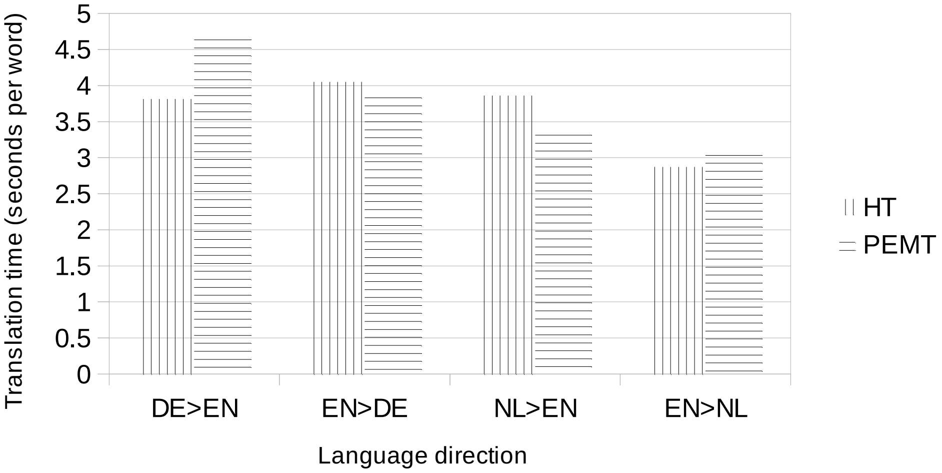

Figure 1. Manual translation and PE time (HT=translation from scratch; PEMT=post-editing).

The results of this part of the objective evaluation are mixed. For English→German and Dutch→English, post-editors took less time than the professionals translating from scratch on average (3.5% and 13.6% productivity gains, respectively). While 13.6% certainly is an encouraging figure, 3.5% represents a modest productivity gain to justify the investment in MT. In real translation workflows, productivity gains of only a few percentage points thanks to PE would be regarded as negative results, in the sense that the management of translation agencies or multilingual departments of large companies would be reluctant to introduce MT with such low gains for a particular language pair; however, it should be noted that our users had no training or experience in PE, and that even relatively marginal productivity gains of this kind would correspond to potentially significant savings across multiple language pairs, such as those typically covered by large multinational companies. In addition, it seems reasonable to expect PE-related productivity gains to rise as staff receive training and acquire experience in the task for a specific language pair. In contrast, for the other two translation directions (i.e. German→English and English→Dutch), post-editing MT output took more time than translating from scratch, leading to productivity losses of 19.16% and 7.88%, respectively, which again can be attributed, at least to some extent, to the fact that the participants in our study had no prior training in PE.

Next, we compared these results with the users' judgements, i.e. their perception regarding both effort and speed, as well as their favourite working method (cf. Tables 1, 2 and 3 for EN—DE, and Tables 4, 5 and 6 for EN—NL). Figure 2 shows these judgements on a 5-point scale.[5] The closer the value of a judgement is to 5, the stronger the preference given to manual translation from scratch. Conversely, the closer the value is to 1, the stronger the preference for post-editing MT output.

Figure 2 also includes the PE gains in terms of time (based on the results presented in Figure 1). The PE time gains are scaled up to a 5-point scale with the following formula:

```
If (PE time gain < 0%)
    Time gain = 3 - 2 * abs(PE time gain)
else
    Time gain = 3 + 2 * 2 * abs(PE time gain)
```

The equation evaluates to 3 (middle value on the 5-point scale, corresponding to no winner between PE and manual translation from scratch) if the PE gain is 0%. The score equals 5 (highest score, i.e. maximum preference for translation from scratch) if PE takes double the time than translating from scratch (i.e. PE productivity gain -50%). Finally, the score equals 1 (lowest score, i.e. maximum advantage for PE) if PE takes half the time compared to translating from scratch (i.e. PE productivity gain 100%).

Analyzing the results shown in Figure 2, we obtain three main findings. First of all, when comparing the subjective judgments with the actual time gains, we notice that most of the judgments are biased towards translation from scratch, with the only exceptions being speed for NL→EN and favourite method for DE→EN. When considering all four translation directions, the scores given to effort, speed and favourite working method are on average 1.07, 0.79 and 1.24 points higher, respectively, than the actual time gain score.

[5] While perceptions of PE effort and speed (as opposed to manual translation from scratch) were originally expressed on a 1-5 scale, overall preference was scored on a 1-3 scale. In the interest of consistency, the values for preference are thus scaled up to a 1-5 scale.

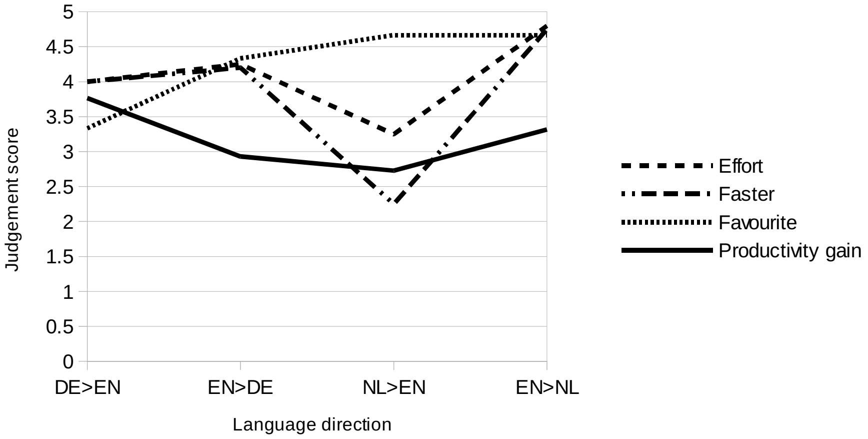

Figure 2. Comparison of users' perceptions and real time gain with PE.

Secondly, all the user judgments tend to express consistent preference for manual translation from scratch over PE of MT output, especially for EN→DE and for EN→NL. The exception to this is NL→EN, where the users perceived PE to be slightly faster (2.25) but requiring more effort than translating from scratch (3.25), but then expressed an overwhelming preference for manual translation as their favourite method (4.67).

Thirdly, while the judgments regarding speed and effort vary across translation directions, probably reflecting the differences in time gain, the results for the favourite working method are rather stable across translation directions, and are particularly high, with similar scores indicating a strong preference for manual translation from scratch, for three translation directions out of four (i.e. all of them except DE→EN). We can thus conclude that the users' overall preference for translation from scratch as their favourite working method is independent of the actual time gain/loss and from real productivity advantages when comparing translation from scratch with MT PE.

5.2 Automatic Evaluation Metrics and PE Gains

The data obtained in this experiment, which consisted of human translations from scratch and of post-edited MT output for the four translation directions, can be considered as alternative sets of reference translations, since the post-edited MT output is of publishable quality; the main difference between the two sets of references is that the translations from scratch were created independently of the MT system, while the post-edited versions were based initially upon the raw statistical MT output, with subsequent revision. One interesting observation in this respect is that, while MT output for a given language pair is consistent (the strengths and weaknesses of a system remain stable, and thus no substantial qualitative variation occurs in the output), human translators as well as post-editors cannot be assumed to be consistent.

This is due only in part to individual differences and idiosyncracies, as two human translators may prefer different, but equally good, translations of the same source passage simply because of personal stylistic preference. One must also consider the inherent variability of human behaviour, including when the same person does a relatively repetitive translation over a period of time: even though they may come across identical phrases at different points (as was likely e.g. with the biographies included in our data sets), they might, more or less consciously, end up translating them differently. This variable behaviour applies even more to post-editors: the degree and the type of corrections made by the same as well as by different individuals to the MT output for one language pair are likely to be unpredictably inconsistent.

We thus evaluated the raw MT output against both the human translations and the post-edited MT output using three state-of-the-art automatic evaluation metrics, namely BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006). The automatic MT evaluation scores are shown in Figure 3. Following the conventions used in Snover et al. (2006), the scores against references translated from scratch are named after the metric (i.e. BLEU, METEOR and TER), while the scores against the post-edited references are named appending the prefix H (i.e. HBLEU, HMETEOR and HTER, respectively).

If we compare the human translation scores against the PE scores, we can see that the PE scores are consistently better than the human translation scores for all the translation directions across all the metrics (bearing in mind that TER is an error rate metric, so unlike the other two metrics, the lower the score the better). This is expected, as the automatic metrics rely on *n*-gram matching of surface forms.[6] Hence, a reference translation that is based on the output of the MT system is likely to have a higher overlap with the raw MT output than a reference that is created independently of the MT output.

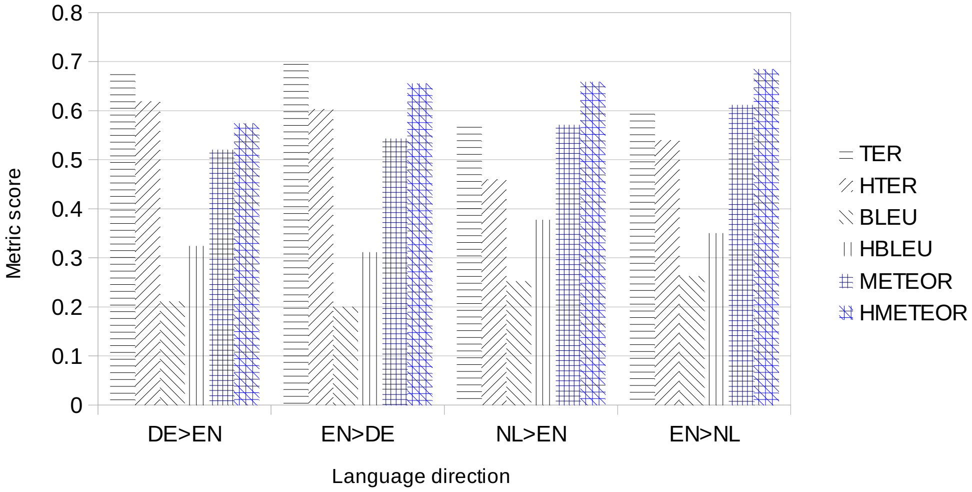

Figure 3. Scores of automatic evaluation metrics.

We then explore whether the scores obtained via the automatic evaluation metrics correlate with translation time, both when translating from scratch and when post-editing. For each translation direction we compute the Pearson correlation between the translation times and the scores of each automatic metric obtained on the MT output at the document level. A limitation that should be taken into account when interpreting any findings extracted from the data relates to the small size of the samples (the number of documents per translation direction varies from 10 to 21).

Intuitively, one would expect translation times to correlate with the scores of the automatic metrics (i.e. the better the score, the lower the PE time). Thus we would expect negative correlations between translation times and BLEU and METEOR scores (i.e. the longer the translation time the worse the metric score) and positive correlations between translation times and TER scores (i.e. the longer it takes to translate, the higher the error rate expressed by TER).

[6] METEOR also considers additional linguistic information, such as stems and synonyms.

It should also be noted that all correlations are calculated with respect to the reference that was translated from scratch. We adopted this approach as these references are independent of the raw statistical MT output, while post-edited references would be biased towards the MT system. Table 7 shows the correlations, which are presented for each translation mode (PEMT, HT), for each metric (TER, BLEU, METEOR) and for each translation direction, plus for all the translation directions combined.[7] The correlations are shown in bold (expected ones), in italics (unexpected ones), and in normal font (no correlation). We consider there to be no correlation if the value is between -0.2 and 0.2.

Translation	*TER*		*BLEU*		*METEOR*	
direction	*PEMT*	*HT*	*PEMT*	*HT*	*PEMT*	*HT*
DE→EN	**0.79**	**0.41**	**-0.93**	**-0.24**	**-0.94**	-0.10
EN→DE	*-0.58*	0.20	*0.53*	**-0.28**	*0.41*	**-0.24**
NL→EN	-0.20	**0.45**	0.00	**-0.38**	**-0.23**	**-0.23**
EN→NL	0.00	-0.05	-0.03	0.19	0.01	0.19
All	**0.25**	**0.28**	**-0.28**	**-0.23**	**-0.42**	**-0.25**

Table 7. Correlations between translation time and automatic metrics.

Considering each of the four translation directions separately, translations from scratch (columns HT) seem to correlate more consistently (out of 12 results, there are 7 expected correlations, 5 no correlations and no unexpected correlations) than post-edited translations (columns PEMT), for which the picture is rather mixed (4 expected correlations, 5 no correlations and 3 unexpected ones). Aggregating the data for all the translation directions, we observe consistent results regardless of the metric (TER, BLEU and METEOR) or the translation method (PEMT, HT): all the correlations are as expected, their values ranging from ±0.23 to ±0.42.

6 Conclusions

We have presented a study of real vs perceived PE productivity gains for the German—English and Dutch—English bidirectional language pairs. Previous studies such as Plitt and Masselot (2010) and Zhechev (2012) had looked at PE productivity gains compared to manual translation. However, in a similar vein to Koehn and Haddow (2009) and Koponen (2012), this study has crucially brought into the picture the perceptions of the users in terms of PE effort and speed, comparing them to the actual PE time gains.

We have found a bias in favour of translation from scratch across all four translation directions for all the levels of perception considered (speed, effort and favourite working method). While the perception of speed and effort seems to correspond to the actual gains to some extent, the favourite working method remains independent of the time gain and is consistently in favour of manual translation from scratch, thus pointing to a lingering sceptical attitude towards the benefits of PE, regardless of actual productivity gains. We have found these for Dutch→English and, albeit more modestly, for English→German, while PE led to productivity losses over manual translation from scratch for English→Dutch and German→English; crucially, PE was consistently the least preferred working method compared to translation from scratch, regardless of the productivity gains or losses.

In addition, we have explored the correlations of three standard automatic evaluation metrics (BLEU, METEOR and TER) with translation time, both when translating manually and when post-editing MT output. Although we can only reach tentative conclusions due to the limited data analyzed across the four language pairs, both manual translation and post-

[7] Note that these correlations are calculated by aggregating the data across all the four translation directions. Thus any findings drawn may be limited due to lack of cohesion of the data.

editing lead to weak correlations between the time to complete the task and the scores of the automatic evaluation metrics.

Acknowledgements

This work was partially funded by the European Commission through the CoSyne project (Grant FP7-ICT-4-248531) and partially supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of CNGL at Dublin City University. The authors would like to thank Tim Koch and Peggy van der Kreeft at Deutsche Welle and Jaap Blom and Johan Oomen at the Netherlands Institute for Sound and Vision, who ran the field tests on which this study was based, and the 20 professionals at both DW and NISV who took part in them. Thanks are also due to the anonymous referees who commented on an earlier version of this paper.

References

Allen, J. (2003). Post-editing. In Somers, H. (ed) *Computers and Translation: A Translator's Guide*. John Benjamins. 297-317.

Banerjee, S. and Lavie, A. (2005). An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL-2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.* Ann Arbor, Michigan. 65-72.

DePalma, D.A. (2011). *The Market for MT Post-Editing in 2011*. Cambridge, MA: Common Sense Advisory. Available at www.commonsenseadvisory.com/AbstractView.aspx?ArticleID=2202.

Guerberof, A. (2009). Productivity and quality in MT post-editing. In *Proceedings of the MT Summit XII Workshop on "Beyond Translation Memories: New Tools for Translators MT"*. Ottawa, Canada. 8 pages.

Koehn, P. and Haddow, B. (2009). Interactive assistance to human translators using statistical machine translation methods. In *Proceedings of the MT Summit XII: Twelfth Machine Translation Summit*. Ottawa, Ontario, Canada. 73-80.

Koponen, M. (2012). Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the WMT 2012 7th Workshop on Statistical Machine Translation*. Montréal, Canada. 181-190.

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

Läubli, S., Fishel, M., Massey, G., Ehrensberger-Dow, M. and Volk, M. (2013). Assessing post-editing efficiency in a realistic translation environment. In *Proceedings of the MT Summit XIV Workshop on Post-editing Technology and Practice*. Nice, France. 83-91.

Martzoukos, S. and Monz, C. (2010). The UvA system description for IWSLT 2010. In *Proceedings of the 7th International Workshop on Spoken Language Translation*. Paris, France. 205-208.

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*. Philadelphia, PA. 311-318.

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

Poulis, A. and Kolovratnik, D. (2012). To post-edit or not to post-edit? Estimating the benefits of MT post-editing for a European organization. In *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice*. San Diego. 60-68.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas "Visions for the Future of Machine Translation"*. Cambridge, MA. 223-231.

Tatsumi, M. and Roturier, J. (2010). Source text characteristics and technical and temporal post-editing effort: what is their relationship? In *Proceedings of the JEC 2010 Second joint EM+/CNGL Workshop "Bringing MT to the user: research on integrating MT in the translation industry"*. Denver, Colorado. 43-51.

Zhechev, V. (2012). Machine translation infrastructure and post-editing performance at Autodesk. In *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice*. San Diego. 87-96.

Zhechev, V. (2014). Analysing the Post-Editing of Machine Translation at Autodesk. In Balling, L.W., Carl, M., Simard, M., Specia, L. and O'Brien, S. (eds) *Post-Editing of Machine Translation: Processes and Applications*. Newcastle upon Tyne: Cambridge Scholars Publishing. 2-23.