JoSTrans
The Journal of Specialised Translation

www.jostrans.org · ISSN: 1740-367X

Sharon O'Brien, R. (2009). Quality and Machine Translation: A realistic objective? *The Journal of Specialised Translation, 11*, 52-74. https://doi.org/10.26034/cm.jostrans.2009.639

Quality and Machine Translation: A realistic objective?

Rebecca Fiederer and Sharon O'Brien, Dublin City University

ABSTRACT

Is Machine Translation (MT) output necessarily of lower quality than human translation? Taking a user guide in English as the source text, German as the target language, and IBM WebSphere as the MT system, we endeavour to answer this question. Eleven suitably qualified raters rated 30 source sentences, three translated versions and three post-edited versions for the parameters *clarity*, *accuracy* and *style*. The results show that the machine translated, post-edited output was judged to be of higher *clarity* and *accuracy*, while the translations were judged to be of better *style*. When asked to pick their "favourite" sentence, the majority of the evaluators chose translated (as opposed to machine translated) sentences. Further, sentences where Controlled Language (CL) rules had been applied scored higher on *clarity* and *accuracy*, adding further to the claim that the application of CL rules improves MT output.

KEYWORDS

Machine translation, machine translation evaluation, translation quality, controlled language.

1. Introduction

In her paper on exploring user acceptance of machine translation output, Bowker cites evidence of a growing demand for translators coupled with a dearth of qualified translation professionals (Bowker and Ehgoetz, 2007: 210). This is one of the reasons for the renewed interest in machine translation (MT), along with other reasons such as the wish to penetrate new markets, the requirement to publish translated material at the same time as source language material and the on-going requirement to reduce the cost of translation. As the interest in, and demand for, MT increases, it is reasonable to assume that translators who work in technical domains will be increasingly required to interact with MT and, yet, research into the topic within Translation Studies is still quite limited. Published research on the topic of MT itself is plentiful but appears at the moment to be focused more on computational and empirical research, such as statistical methods in MT and automatic evaluation of output (see, for example, the proceedings of the recent MT Summit XI [Maegaard 2007] and the proceedings of TMI 2007 [Way and Gawronska 2007]). It is important for translators to both keep abreast of developments in this area and to actively contribute to it so that the translation community can engage with technological demands and provide well-informed feedback to system developers, end users and translation customers.

If MT is used for 'gisting' purposes, then often no post-editing is required unless, of course, the user would like to see a more polished version of the output. However, when MT is used for publication purposes, then some level of post-editing is normally required. Another method for improving MT output is to apply Controlled Language (CL) rules to the source text in order to reduce ambiguities and complexity. CL rules generally make the source text input more suitable for MT by reducing sentence length, eliminating problematic features such as gerunds, long noun phrases, ambiguous anaphoric referents and so on. Such features are often termed "Translatability Indicators" (Underwood and Jongejan 2001) or, more precisely, "Negative Translatability Indicators" (NTIs for short) (O'Brien 2005). It has been shown by several researchers that the use of CL rules can have a positive impact on MT output (Aikawa et al 2007; Bernth 1998; Mitamura and Nyberg 1995; Mitamura 1999). The influence of CL on the quality of MT output is an important topic in this paper and will be discussed further in the Data Analysis section.

In a previous study by one of the authors of this paper using an English software user manual translated into German by the IBM WebSphere MT engine (O'Brien 2006), the effect of CL rules on temporal, technical and cognitive post-editing effort was measured (cf. Krings 2001) and the findings were that post-editing effort can be reduced by removing NTIs from the source text. The removal of NTIs involved making sure that the selected sentences did not contain any of 29 selected NTIs (e.g. sentences over 25 words, passive voice etc), and that was the extent of the editing on the selected sentences. In that study, it was also found that the removal of some NTIs had a greater impact on post-editing effort than the removal of others. One question which is close to the heart of language professionals remained unanswered by this study: how does the quality of the post-edited product compare with the quality of human translation? We assume that many language professionals would predict that the product of machine translation combined with post-editing is inferior in quality to the human translated product. This paper investigates that assumption in both a qualitative and quantitative manner by conducting a comparative evaluation of quality for sentences produced by MT and subsequently post-edited and sentences that have been translated by humans. In both scenarios, the work was completed by experienced professional translators who were paid for their time.

Before discussing our methodology and results, we will first give some consideration to the topic of MT evaluation. As this is a broad topic, we will restrict our discussion to what we feel is relevant for the study described here.

1.1 MT Evaluation

Much has been written about the evaluation of MT output (e.g. Hutchins and Somers 1992; Arnold et al 1994; White 2003; King, Popescu-Belis and Hovy 2003; Coughlin 2003; Babych and Hartley 2004; Hamon et al 2007). In the early years of MT evaluation, human evaluators were necessarily involved in the exercise. However, the use of human judges invariably brings with it a number of issues, not least of which are subjectivity, cost and time. In an effort to eliminate these issues, automated evaluation metrics have been developed such as *BLEU, NIST* and *WNMf* (weighted N-Gram model) (Papineni et al 2001; Doddington 2002; Babych et al 2004). The main presupposition behind BLEU, for example, is that the closer MT output is to a professional human translation, the better it is. This *closeness* is measured numerically. On the one hand, it has been demonstrated that such automated metrics correlate well with human judgments (Hamon et al 2007) and yet other researchers have claimed that perceived correlations may not be as high as previously thought (Callison-Burch et al 2006). As researchers work with these metrics, we may well see a development in the not-too-distant future where MT systems will only be evaluated by automated metrics. Somewhat ironically, measures such as BLEU still require human input because the metric compares MT output with so-called 'gold standard' *human* translations.

Most research on MT evaluation is concerned with evaluating raw MT output rather than post-edited text, as is the case in our study. It is our contention that a comparison of raw MT output with the final version of a human translation is an unequal comparison. Rather, post-edited MT output ought to be the basis for comparison with human versions. According to Hutchins and Somers (1992: 163) the most obvious tests of the quality of a translation are:

- *Fidelity* or *accuracy*: the extent to which the translated text contains the "same" information as the original
- *Intelligibility* or *clarity*: the ease with which a reader can understand the translation
- *Style*: the extent to which the translation uses the language appropriate to its content and intention.

Arnold et al (1994: 169) also state that *intelligibility* is "a traditional way of assessing the quality of translation". This *intelligibility* may be affected by grammatical errors, mistranslations and untranslated words. They list *accuracy* as another important factor in checking whether the meaning of the source language sentence is preserved in the translation (ibid 1994: 171). Arnold et al claim that high *intelligibility* normally means high *accuracy*, as these scores are often closely related. Dabbadie et al (2002: 14) also state that *intelligibility* is one of the most frequently used metrics of the quality of output, and Roturier (2006: 83) points out that two of the most frequently used human evaluation metrics of MT output quality are *intelligibility* and *fidelity*. T.C Halliday, in Van Slype's 1979 study on evaluating MT quality for the European Commission (cited in Dabbadie et al 2002: 14), proposed a measurement of *intelligibility* on a 4-point scale. For Halliday, *comprehensibility* and *intelligibility* are synonymous. In Halliday's scale, outlined here, *intelligibility* is rated from 0 through to 3, with 3 being the most intelligible:

> 3 – Very intelligible: all the content of the message is comprehensible, even if there are errors of style and/or of spelling, and if certain words are missing, or are badly translated, but close to the target language.
>
> 2 – Fairly intelligible: the major part of the message passes.
>
> 1 – Barely intelligible: a part only of the content is understandable, representing less than 50% of the message.
>
> 0 – Unintelligible: nothing or almost nothing of the message is comprehensible.

As can be ascertained, before one even broaches the problem of subjectivity, MT evaluation is hampered by the use of synonyms to describe evaluation parameters. The creation of FEMTI (the Framework for the Evaluation of Machine Translation in ISLE), through the International Standards for Language Engineering (ISLE) project, was an attempt to gather into one place the accumulated experience of MT evaluation (King, Popescu-Belis and Hovy 2003). ISLE proposed the measures *comprehensibility, readability, style* and *clarity*, which Vanni and Miller (2002 cited in Dabbadie et al 2002: 14) merged into one single evaluation feature: *clarity*. Vanni and Miller's *clarity* measure ranges from between 0 (meaning of sentence is not apparent, even after some reflection) and 3 (meaning of sentence is perfectly clear on first reading). Arnold et al (1994: 170) point out that *intelligibility* (or *clarity*) might seem a useful measure of translation quality, as scoring for *intelligibility* reflects directly the quality judgment of the user: the less the user understands, the lower the score. Hutchins and Somers (1992: 164) further state that, generally, only individual sentences are evaluated – making *intelligibility* tests even more subjective and uncertain than they might be if whole texts were rated.

With regards to *fidelity*, various tests have been proposed and implemented (Hutchins and Somers 1992: 163). For the ALPAC report, evaluators were asked to read MT output and then judge how much more 'informative' the original was. This can obviously be criticised as being "excessively subjective" (ibid). Hutchins and Somers state that more objective tests of *fidelity* for instruction manuals, for example, would be a practical performance evaluation. For example, the question "Can someone using the translation carry out the instructions as well as someone using the original?" may be asked of evaluators (ibid). Bowker and Ehgoetz (2007) reflect on the appropriateness of a *recipient evaluation*, i.e. an evaluation by recipients of the target text, which not only takes quality into account, but also cost and speed of the translation.

Arnold et al (1994: 169) state, while talking about *intelligibility*, that some studies take *style* into account, although it "does not really affect the *intelligibility* of a sentence". Hutchins and Somers (1992: 164) maintain that measurements of *style* are "as subjective as the global rankings of *intelligibility*". However, the appropriateness of a particular *style* still remains an "important factor" (ibid). Style, while being considered important in the rating of human translation, in particular in literary domains, rarely figures as an evaluation parameter in the rating of MT output. Consideration was given to the different evaluation parameters before deciding on a framework for this study and the selected parameters are discussed in the next section on Methodology.

2. Methodology

2.1 Evaluation Criteria

For this study, three evaluation criteria were chosen:

Clarity
Accuracy
Style

As *clarity* and *intelligibility* are synonymous in much of the literature, *clarity* was chosen as the title for the first parameter as, in our opinion, the term *clarity* could be easier for evaluators to understand than *intelligibility*. *Clarity* is an obvious and fundamental criterion for evaluating translation quality, and therefore constitutes the first parameter in our evaluation framework. Evaluators were guided in their interpretation of *clarity* by the provision of the following question: *How easily can you understand the translation?* The evaluators then had to answer this on a scale of 1-4, where 1 meant "not understandable" and 4 meant "fully understandable" as shown in Figure 1.

Obviously, translated text might be easily understood, but it may not be an accurate representation of the source text. Therefore, the second criterion chosen was *accuracy*, sometimes used synonymously with *fidelity*. We opted to use the term *accuracy*, simply because it would be confusing for evaluators to be confronted with two terms. According to Arnold et al (1994: 171), scoring for *accuracy* is normally done in combination with (but after) scoring for *intelligibility*. As much of the literature on MT evaluation includes *accuracy* as a parameter, it forms a vital part of this research too. As with the parameter *clarity*, evaluators were given guidance on their interpretation of this parameter through the question and possible answers (see Figure 1): *To what extent does the translation contain the "same" information as the source text?* And *If the sentence contains instructions, do you think someone using the translation could carry out the instructions as well as someone using the original?*

The third and last chosen criterion was *style*. As already mentioned, style rarely occurs as a parameter in MT evaluations. Given that our aim was to compare human translation quality with machine translation and post-editing quality, we felt that the inclusion of style as a parameter was justified. Also, in O'Brien's study (2006) the post-editors (all professional translators at IBM) were advised that "it is not necessary to change text that is accurate and acceptable just for the sake of improving its style", but Hutchins and Somers (1992: 173) point out that:

> Translators are naturally reluctant to be responsible for what they consider an inferior product. Their instinct is to revise MT output to a quality expected from human translators, and they are as concerned with 'stylistic' quality as with accuracy and intelligibility.

This provided us with further impetus to include style in the evaluation framework. Again, the evaluators were guided by some questions: *Is the language used appropriate for a software product user manual? Does it sound natural and idiomatic? Does it flow well?* (Figure 1).

Our assumption was that translators would score higher than post-editors for all three parameters, and we hypothesised that translators would score markedly higher for the parameter *style* because, we assumed, post-editors may be subject to the seductiveness of the MT output. In other words, post-editors may be led to accept what the computer offers.

Scoring scales in past studies have ranged from nine-point-scales (the ALPAC report) to scales of just two (e.g. intelligible, unintelligible). Arnold et al (1994: 170) propose a four-point scale as being more appropriate, and following this guideline a four-point scale was devised for all three parameters in this study. Figure 1 shows the criteria evaluators were provided with in detail:

You will be asked to judge each sentence on a scale of 1 – 4 (1 being the **lowest** score and 4 being the **highest**) for each of the following criteria:

1. CLARITY

• ***How easily can you understand the translation?***

1 – Not understandable.

2 – Only small part understandable.

3 – Mostly understandable.

4 – Fully understandable.

2. ACCURACY

• ***To what extent does the translation contain the 'same' information as the source text? If the sentence contains instructions, do you think someone using the translation could carry out the instructions as well as someone using the original?***

1 – Not the same information. Instructions could not be carried out.

2 – Only some information is the same. Instructions could not be carried out very well.

3 – Most of the same information. Instructions could be carried out nearly as well.

4 – Same information. Instructions could be carried out just as well.

3. STYLE
• ***Is the language used appropriate for a software product user manual? Does it sound natural and idiomatic? Does it flow well?***

1 – Language is inappropriate. Not natural and idiomatic; does not flow well.

2 – Most of the language is inappropriate. Not very natural and idiomatic; does not flow very well.

3 – Most of the language is appropriate. Mostly natural and idiomatic; flows fairly well.

4 – Language is appropriate. Completely natural and idiomatic; flows very well.

Figure 1: Criteria for assessment on which evaluators based ratings

2.2 Data Selection

For the current research, 130 source text sentences, extracted from one section of a software user manual, were available from O'Brien's 2006 study, complete with nine machine-translated and post-edited versions and three versions produced by human translators for each sentence.[1] As this would have generated a large amount of data to be manually evaluated (1,560 sentences in total), and also because time restrictions applied to the study, a decision had to be made regarding how much data and which parts were to be used. A decision was made to use 30 source text sentences, 15 of which contained NTIs as we were keen to examine any differences in quality scores for post-editors and translators when CL rules had been applied. The average length of the sentences was 13 words, with the longest sentence amounting to 23 words and the shortest amounting to 6. The sentences were primarily descriptive, i.e. they described a feature of the particular software program about which the manual was written. Seven of the sentences could be classified as procedural, i.e. they instructed the user to perform an action such as clicking on a particular icon or menu item (see Appendix A for a small sample of the source sentences, translations and post-edited versions).

The data from three post-editors out of the nine were selected so that there was an even number of post-editors and translators (3 + 3). As we did not want to skew the results either for or against post-editing or translation quality, we did not deliberately select either the worst or best post-editors. Therefore, we simply selected the first three post-editors in the list.

The 30 ST sentences were selected chronologically, unless there was a problem such as a missing translation by one of the translators or post-editors. The following NTIs were included in 15 of the 30 sentences from O'Brien's 2006 study:

- Slang *(two instances)*
- Missing relative pronoun
- Proper noun *(nine instances)*
- Abbreviation *(two instances)*
- Gerund *(six instances)*
- Use of "and/or" *(two instances)*
- Ambiguous coordination due to ellipsis
- Post-modifying adjectival phrase
- "(s)" for plural
- Ungrammatical sentence
- Missing relative pronoun + finite verb "which is" or "that is"
- Potentially problematic punctuation
- Ambiguous non-finite verb
- Misspelling
- Personal Pronoun *(two instances)*
- Not an independent syntactic unit *(two instances)*
- Ambiguous non-finite verb phrase
- Problematic punctuation
- Missing "in order (to)"

Thirty sentences is admittedly a small sample, but the total number of sentences to be read and evaluated by each evaluator was 180 (30 source text sentences x 6 versions) and evaluators had to judge each sentence according to three parameters. Since this type of evaluation can be tedious, we felt that the sample could not be made bigger without introducing the risk of boredom and its negative consequences such as inconsistent evaluation or diminishing powers of judgement. When dealing with human evaluation, there is a necessary trade-off between the size of the sample being evaluated and the integrity of the results. This is, of course, one of the weaknesses in human evaluation as a methodology and one of the driving forces in the search for reliable automated metrics, as discussed earlier.

2.3 Additional Parameters

In an attempt to see if evaluators had a natural preference for human translated sentences over machine-translated and post-edited sentences, evaluators were asked to select their favourite translation out of the six given for each sentence.

As well as the criteria for assessment, evaluators were given a briefing and a sample evaluation. In the briefing, evaluators were advised that they would anonymously rate six translations for each of the 30 ST sentences. They were not told that half of the translations were produced by translators and half the translations were in fact post-edited MT output. Arnold et al (1994: 8) point out that a common misconception is that "MT threatens the jobs of translators". Hutchins and Somers (1992: 173) further state that judgement of MT output is often clouded by translators' attitudes and Schäler (1998: 153) indicates that translators widely agree that "MT does not work". Roturier (2006: 81) also highlights the problem that professional translators could be biased when evaluating MT output, as they may perceive MT as a threat. This factor may consciously or unconsciously influence the way translators score MT output (or, in this case, post-edited MT). For these reasons, evaluators were not told of the involvement of MT output in the evaluation exercise. Sentences were presented to evaluators in a randomly mixed sequence with a post-edited sentence first, followed by two translated sentences, two post-edited sentences and one translated sentence. Finally, the evaluation framework was carried out on-screen in Microsoft Excel format.

For O'Brien's study, post-editors were briefed to perform a full post-edit on the German target text so that it met the following criteria:

- Any non-sensical sentences or phrases are repaired
- Any inaccuracies in the information are fixed
- Any mis-translation, non-translation or inconsistent translation of terminology is rectified
- The text is understandable and stylistically acceptable to a German native speaker who needs to understand the contents of the document

It is important to point out here that post-editors were advised not to change text that was accurate and acceptable just for the sake of improving its style.

2.4 Evaluators' Profile

A total number of 15 evaluators agreed to take part in this study. Seven of these had just completed the MA in Translation Studies at Dublin City University (DCU). All evaluators fulfilled the following criteria:

- They were native speakers of German.
- They held an MA in Translation or equivalent translation degree at post-graduate level.
- They had a high level of competence in English, and had all studied abroad in an English-speaking environment.

We were aware of the fact that evaluators were in the process of becoming professional translators and that this may have influenced the results, as they may have judged translation quality harsher than a layperson or end-user would. Arnold et al. (1994: 180), for example, point out that the experienced translator/post-editor is more critical towards translation quality than the customer is. Roturier's (2006) findings support this argument, as he concluded that there was a lack of agreement between translators' judgment of translation quality of web-based information and end-users' judgment. Roturier (2006: 201) states that:

> Users and translators do not have the same linguistic standards and expectations, so translators may be more likely to be affected by translation inaccuracies than users, especially when these translation inaccuracies are generated by MT systems.

To add to this evidence, Bowker and Ehgoetz (2007) point to a similar issue in their paper describing a user evaluation of MT, i.e. when language quality is assessed by evaluators who are highly sensitised to linguistic quality, the results are likely to be more critical than when less sensitised evaluators are used.

The duration of the evaluation was estimated to be approximately one hour, as a pilot run took 45 minutes. Evaluators were advised to carry out the evaluation on-screen and in one sitting, with a break after the first 30 minutes to reduce the effects of fatigue or boredom.

3. Data Analysis

A total of 11 out of the 15 evaluators returned their completed evaluations in the given time frame. Those who did not complete the evaluation offered personal reasons such as a lack of time as an explanation. Although it was disappointing that not all evaluations were completed, it still amounted to a reasonable number of scorers. Whilst discussing MT evaluation, Dyson and Hannah (1987: 166) stress that the number of evaluators should be no fewer than three. Arnold et al. (1994: 171) point out that a reasonably sized group of evaluators/scorers must be used to score MT output. They suggest four scorers as the minimum, but state that a bigger group would make the results more reliable. We were satisfied that a group of 11 evaluators would give us a reliable judgement of post-edited versus human translated quality, at least within the limitations of this study.

Using Microsoft Excel, the results were then analysed. Averages were calculated for the comprehensive set of data. As mentioned previously, evaluators scored 180 sentences for the parameters *clarity, accuracy* and *style*, and marked their favourite translation out of the six given for each of the 30 segments. Separate averages were calculated for the 15 sentences containing Negative Translatability Indicators (NTIs) and sentences where NTIs had been removed. The Wilcoxon test was used to determine whether results have statistical significance. The null-hypothesis was that there would be no significant difference in scores for translators and post-editors. The null hypothesis can be rejected when the probability value (p-value) is less than 0.05.

3.1 Results – Clarity

The first parameter evaluators were asked to score was *clarity*. For all 30 sentences combined, there was only a minimal difference between translators' and post-editors' scores. All three translators combined achieved an average score of **3.44** out of a highest possible rating of **4**. post-editors meanwhile achieved a combined average score of **3.43**. The Wilcoxon signed-rank test (p=.782) shows that there is no significant difference in the results for this parameter. We can therefore say that evaluators found translators' and post-editors' output equally understandable. The score both translators and post-editors were closest to was 3, which was defined as "*Mostly understandable"* as answer to the question "*How easily can you understand the translation?*" in the criteria for assessment (see *Figure 1*).

3.2 Results – Accuracy

Accuracy was the second parameter evaluators were asked to mark. Overall, post-editors scored higher than translators: an average score of **3.47** out of a highest possible score of **4.** Translators meanwhile achieved a combined average score of **3.32**. The Wilcoxon test (p=.0001) for this parameter reveals that there is indeed a significant difference between translators' and post-editors' scores for this parameter. According to these ratings, evaluators deemed post-editors' output to be more accurate than that produced by translators.

According to Krings (2001: 7), post-editors focus on adjusting the MT output so that it reflects as *accurately* as possible the meaning of the source text. Post-editors may have scored higher as they were potentially more focused on accurately reflecting ST meaning than translators, who may have been more focused on the stylistic quality of the target text, for example. In relation to the criteria for assessment, this means both translators and post-editors were closest to the score of 3, which answered the question "*To what extent does the translation contain the 'same' information as the source text? If the sentence contains instructions, do you think someone using the translation could carry out the instructions as well as someone using the original?"* with the answer "*Most of the same information. Instructions could be carried out nearly as well"*.

3.3 Results – Style

The third parameter evaluators were asked to score was *style*. This is the only category where, overall, translators scored higher than post-editors. Translators achieved a combined average of **3.23** out of the highest rating of **4**. Post-editors meanwhile achieved an average of **3.03** – the lowest combined average score out of the three parameters. The Wilcoxon test (p=.0001) for this parameter reveals that there is a significant difference between translators' and post-editors' scores. Evaluators rated the *style* of translators' output higher than that of post-editors' output.

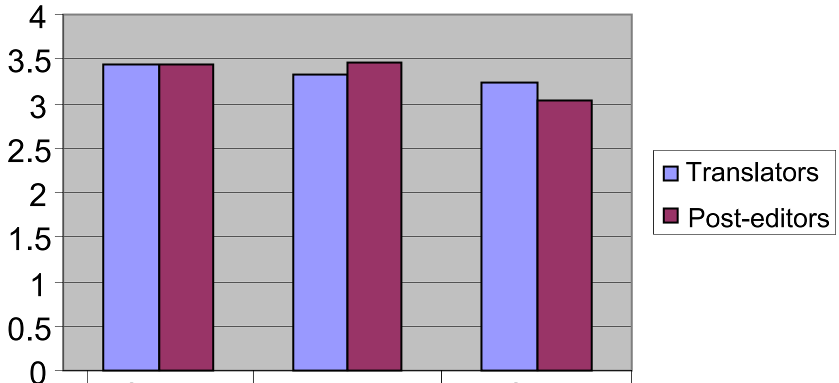
Out of the three parameters, *style* shows the biggest difference between translators' and post-editors' scores (0.20). However, both translators and post-editors scored lower for this parameter than for *clarity* and *accuracy*. This could be due to the fact that the judgement of *style* is somewhat more subjective than for the other two parameters and, as we have already discussed, professional linguists may be more critical of *style*.

It is also important to once again highlight here that during the generation of the data in O'Brien's 2006 study, post-editors were told in the briefing that "it is not necessary to change text that is accurate and acceptable just for the sake of improving its style". As Senez (1998, pages not numbered) points out: A translator will always strive to disguise the fact that the text has been translated. In the case of post-editing, it is enough for the text to conform to the basic rules of the target language, even if it closely follows the source text.

Therefore, it can be concluded that post-editors might have been more concerned with the criteria *clarity* and *accuracy*, while translators may have also focused on the *style* of the translation. With regards to the criteria for assessment, both translators and post-editors are closest to the score of 3, which rates style as *"Most of the language is appropriate. Mostly natural and idiomatic; flows fairly well".*

3.4 Results – Summary of All Parameters

Figure 2 provides a summary of the results for all parameters together.



	Clarity	Accuracy	Style
□ Translators	3.44	3.32	3.23
■ Post-editors	3.43	3.47	3.03

Figure 2: Average Scores for Translators and Post-editors

A difference can be observed in the evaluators' selection of their favourite target output for each of the 30 sentences. 63% of favourite sentences was produced by translators, while only 37% was produced by post-editors as is shown in Figure 3. The possible reasons for this will be discussed further below.

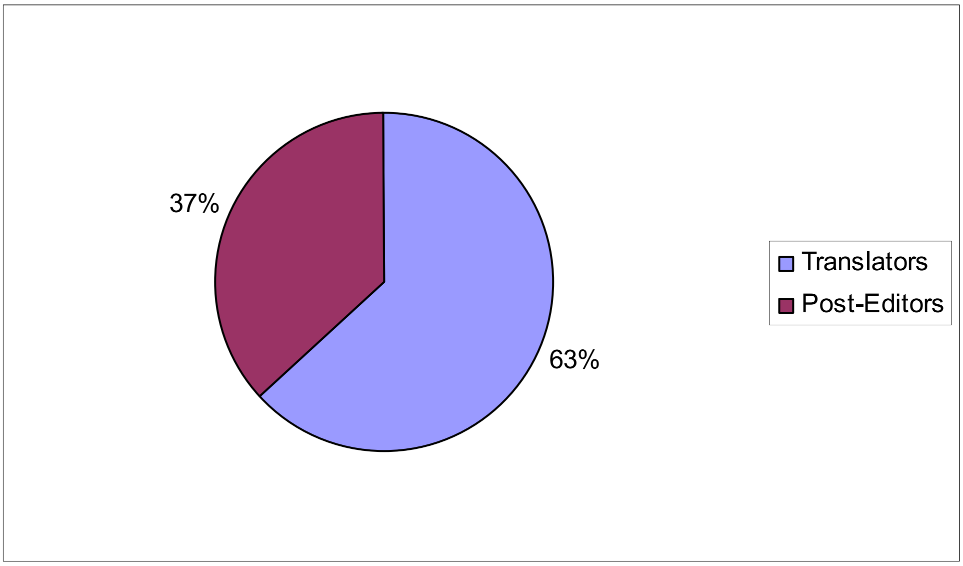
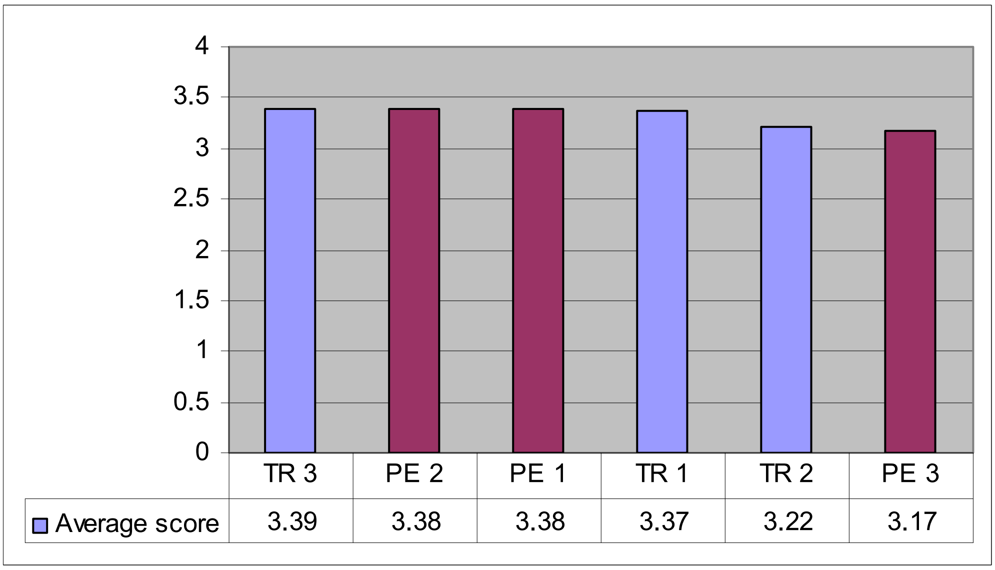


Figure 3: Evaluators' Favourite Sentences

3.5 Individual Scores: Post-editors versus Translators

Taking into account all three parameters, translator 3 achieved the highest average score from evaluators. Figure 4 shows the full rankings from highest-ranked (translator 3) to lowest-ranked (post-editor 3).



	TR 3	PE 2	PE 1	TR 1	TR 2	PE 3
□ Average score	3.39	3.38	3.38	3.37	3.22	3.17

Figure 4: Translators' and Post-editors' Average Scores and Rankings for all Parameters

In general, scores between the first four places (translator 3, post-editor 2, post-editor 1 and translator 1) are very similar. Post-editor 2 and post-editor 1 share second position with an average score of 3.38, only a 0.01 difference to translator 3 (3.39). Translator 1 is in fourth position with an average score of 3.37. Post-editor 3 achieved the lowest overall average score.

More marked changes can be observed when looking at the rankings for each of the parameters *clarity, accuracy* and *style*. When measuring scores for *clarity* (as seen in Figure 5), translator 3 remains in first position, followed by post-editor 1.

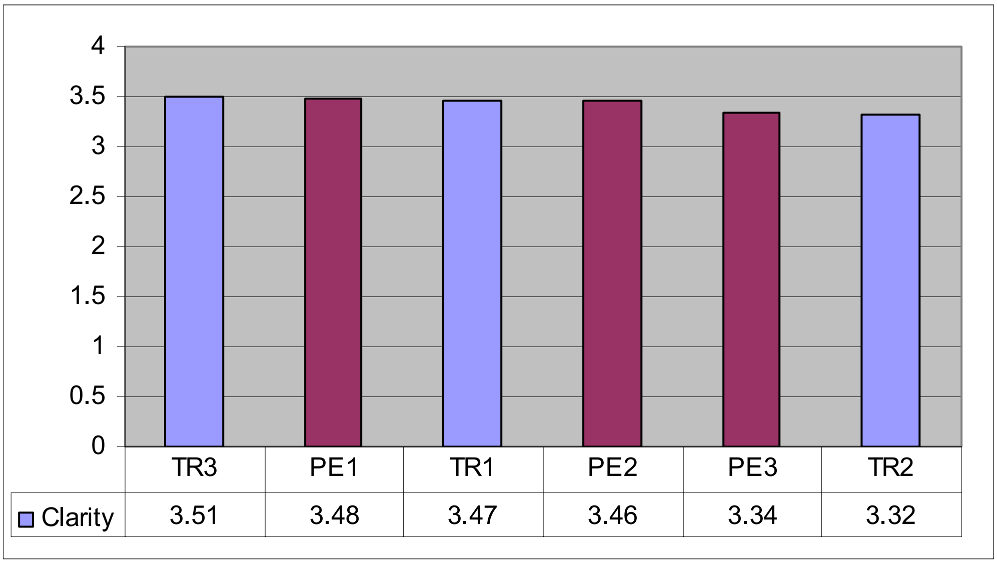


Figure 5: Translators' and Post-editors' Average Scores and Rankings for Clarity

As Figure 6 shows, post-editor 1 and post-editor 2 achieved the highest scores for the parameter *accuracy*. This is the category where translators scored the lowest marks, as is evident in Figure 6. Particularly translator 2 achieved a low score (3.1) compared to the second-to-last positioned post-editor 3 (3.32).

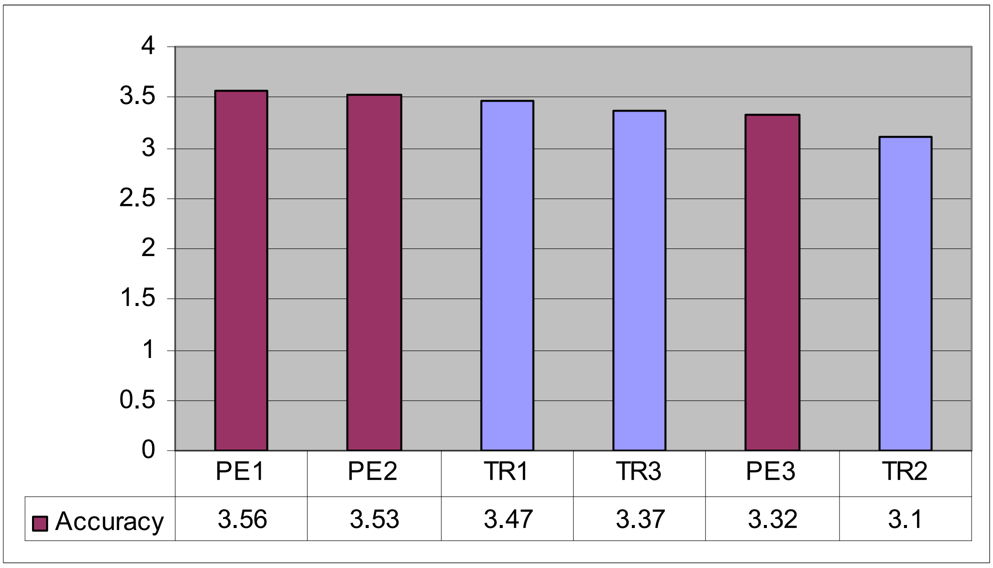


Figure 6: Translators' and Post-editors' Average Scores and Rankings for Accuracy

In the final category *style*, a marked change in rankings can be observed, as Figure 7 shows. The three translators were ranked highest in this category.

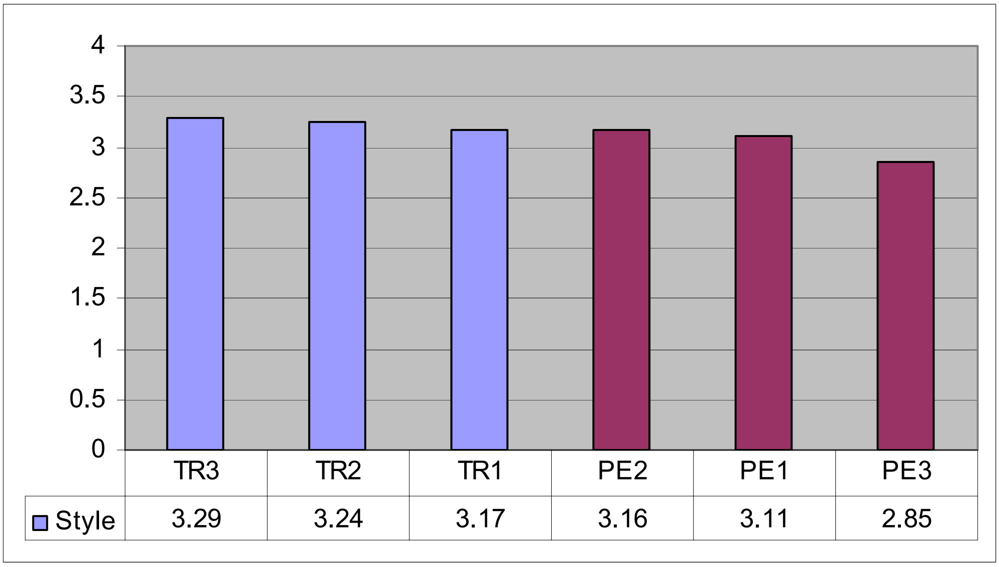


Figure 7: Translators' and Post-editors' Average Scores and Rankings for Style

3.6 Favourite sentences

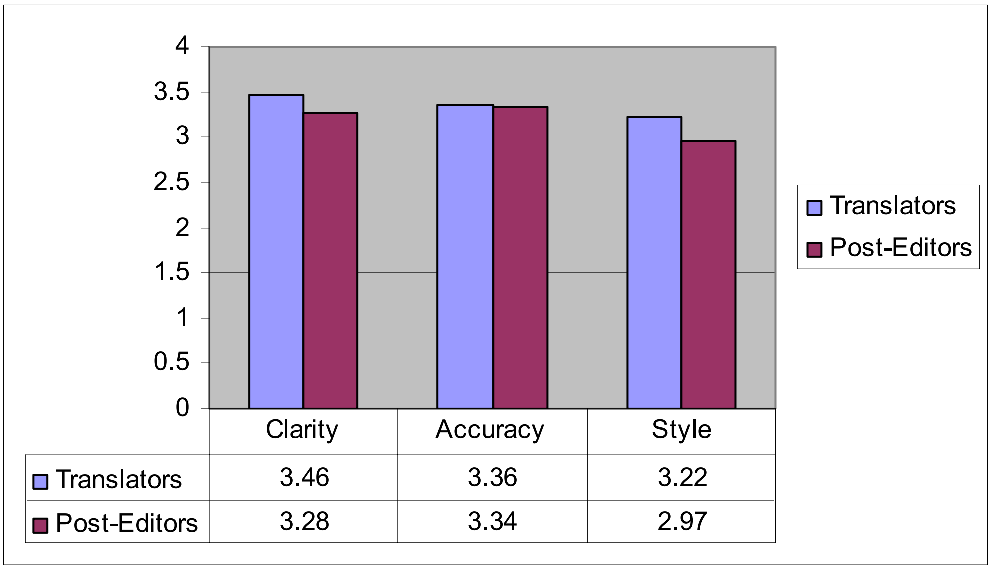
Evaluators showed a clear preference for translators' sentences over post-editors' sentences: 63% to 37%. This is despite the fact that post-editors overall scored significantly higher for *accuracy* and marginally higher for *clarity*. This finding indicates that evaluators preferred translated over post-edited sentences for reasons other than the criteria given in this study, or that they considered *style* to carry more weight than *clarity* and *accuracy*.

Out of the 11 evaluators, ten preferred translators' over post-editors' sentences overall (the greatest difference being 23 translated sentences to seven post-edited sentences; the smallest difference being 14 translated sentences to 12 post-edited sentences). Only one evaluator (Subject F) marked more post-edited sentences as their favourite, choosing 16 post-edited sentences and 14 translated sentences.

In some instances, evaluators picked one sentence as their favourite despite not giving this sentence the highest marks, or giving other sentences exactly the same marks in all three categories. For example, Subject C chose translator 2's translation of ST3 as their favourite, despite giving post-editor 2 the same scores for all three categories. The reason given by Subject C was that their favourite sentence by translator 2 was the "most idiomatic". Subject I chose translator 1 and translator 3's translations (which were identical) as their favourite for ST9, stating: "Sehr schön, der Ausgangstext ist sehr schlecht formuliert, diese ÜS erheblich besser" (Very nice, the source text is very badly phrased, this translation is much better)^{2}. Subject J also chose translator 1 and translator 3’s translation for this source text, stating: “problem: it's not clear in the original what ‘mentioned’ refers to; but this translation makes sense”. Subject J selected translator 2’s translation as their favourite for ST12, stating it is “nice, short and to the point”. Only translator 2 scored 4 marks in all three categories in this instance.

3.7 Impact of NTIs

The presence of NTIs (where CL rules were not applied to the source text) had a marked impact on scores given by evaluators. For the 15 sentences containing NTIs such as slang, proper nouns, gerunds, abbreviations or ungrammatical sentences, translators scored higher than post-editors for all three parameters. This is illustrated in Figure 8.



	Clarity	Accuracy	Style
■ Translators	3.46	3.36	3.22
■ Post-Editors	3.28	3.34	2.97

Figure 8: Translators’ and Post-editors’ Average Scores for Sentences containing NTIs

Results of the Wilcoxon test here reveal a statistically significant difference in scores for *clarity* (p=.0001) and *style* (p=.0001). However, the difference in scores for *accuracy* is not statistically significant (p=.675).

Meanwhile for the 15 sentences containing minimal NTIs, i.e. where CL rules had been applied to MT input and all known NTIs had been removed, roles were reversed and post-editors scored higher for *clarity* and *accuracy*. However, translators still scored higher for the parameter *style*, as Figure 9 demonstrates.

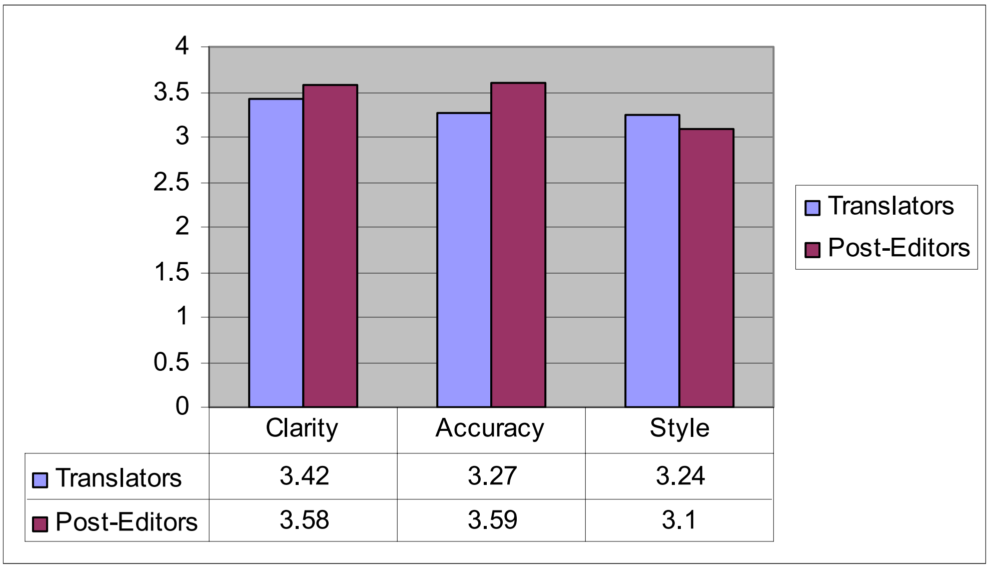


Figure 9: Translators' and Post-editors' Average Scores for Sentences containing Minimal NTIs

The results of the Wilcoxon test for sentences containing minimal NTIs show that there is a significant difference in results for post-editors' and translators' scores for all three parameters: *clarity* (p=.0001), *accuracy* (p=.0001) and *style* (p=.001).

As was hypothesised, the results indicate that the presence of NTIs negatively influenced post-editors' performance, as they scored higher in sentences where NTIs were minimal, i.e. where CL rules were applied to MT input. This can be attributed to the fact that NTIs can cause problems for MT systems; as a result, the MT output post-editors were working with was likely of poorer quality if CL rules had not been applied to MT input. As we outlined in our introduction, many researchers state that implementing CL has a marked effect on MT output. Therefore, it is not surprising that post-editors' performance suffered when all known NTIs were not removed from MT input.

This finding is also in keeping with the findings in O'Brien's study, i.e. that the application of CL rules to a source text prior to MT can reduce post-editing effort, presumably because the output is then clearer and more accurate than when CL rules are not applied.

4. Summary and Conclusions

Our study focused on the question: how does the quality of the post-edited product compare with the quality of human translation? We assumed, on the basis of attitudes in general to machine translation, that most language professionals would predict that the post-edited quality would be inferior to the quality of human translation. However, when quality is defined along the parameters of clarity and accuracy, our results are that the post-edited quality is on a par, if not greater than, human translation. Yet, when style is included as a parameter of quality, then human translation is preferred over the post-edited product. We also found that, when evaluators were asked to select their "favourite" sentence, there was a clear preference for human translation. In addition, when we investigated the influence of CL rules, we noted that quality was deemed to be higher when CL rules were applied to the source sentence.

Clearly, this study is limited in terms of language pair, direction, text type, MT engine and size. Despite this, we feel that it is appropriate to draw some conclusions here, keeping in mind that a specialised text type and one language pair were the focus of the study. Our first conclusion is that machine translation plus post-editing does not necessarily produce a product of inferior quality. In fact, our data suggest that the post-edited quality can be equal to or higher than human translation quality when clarity and accuracy are taken into account. This is presumably influenced by the quality of the raw machine translation output which, if reasonably good to begin with, gives the post-editor a better chance of polishing the text quickly. Secondly, it would appear that evaluators prefer the *style* (as defined in our study) of human translation. However, we must take into account that users of MT are less likely to be professional linguists, and more likely to be users of information; for example, in the IT domain, they may simply need to know how to install a product, use it, etc. Many end users may not care much about *style* as long as they can access and understand the information they need. Researchers in the domain of MT evaluation (Roturier 2006; Bowker and Ehgoetz 2007) point to the fact that end users' acceptance of MT output may be higher than that of professional linguists. More research into end users' acceptance of MT output and into the implications that might have for our notions of "linguistic quality" is required. Finally, we have seen that there may be correlations between the use of CL rules and the quality of the post-edited product. This is in keeping with previous studies that demonstrate the advantages of using CL rules.

The trend seems to be moving in the direction of increased usage of MT and evidence shows that, when used intelligently, MT does not have to be synonymous with poor quality translation.

References

- Aikawa, Takako et al (2007). "Impact of Controlled Language on Translation Quality and Post-Editing in a Statistical Machine Translation Environment". In *Proceedings of the MT Summit XI*. Copenhagen, Denmark, 10-14 September. 1-7.

- Arnold, Doug et al (1994). *Machine Translation: An Introductory Guide*. Oxford: NCC Blackwell.

- Babych, Bogdan, Elliott, Debbie and Hartley, Anthony (2004). "Extending MT Evaluation Tools with Translation Complexity Metrics". In *Proceedings of Coling 2004 (20th International Conference on Computational Linguistics)*. Geneva, Switzerland, August 2004, 106-112.

- Babych, Bogdan and Hartley, Anthony. (2004). "Extending the BLEU MT Evaluation Metric with Frequency Ratings". In *Proceedings of ACL 2004 (42nd Annual Meeting of the Association of Computational Linguistics)*. Barcelona, Spain, 621-628.

- Bernth, Arendse (1998). "Easy English: Preprocessing for MT" in *Proceedings of CLAW 98 (the Second International Workshop on Controlled Language Applications)*. Pittsburgh, Pennsylvania: Language Technology Institute, Carnegie Mellon University, 30-41.

- Bowker, Lynne and Ehgoetz, Melissa (2007). "Exploring User Acceptance of Machine Translation Output: A Recipient Evaluation". Dorothy Kenny and Kyongjoo Ryou (Eds) (2007). *Across Boundaries: International Perspectives on Translation*. Newcastle-upon-Tyne: Cambridge Scholars Publishing, 209-224.

- Callison-Burch, Chris, Osborne, Miles, Koehn, Philipp (2006). "Re-evaluating the role of BLEU in Machine Translation Research". In *Proceedings of EACL 2006 (11th Conference of the European Chapter of the Association of Computational Linguistics)*. Trento, Italy, 249-246.

- Coughlin, Deborah (2003). "Correlating Automated and Human Assessments of Machine Translation Evaluation". In *Online Proceedings of the MT Summit IX (2003)*. New Orleans. Online at:
http://www.amtaweb.org/summit/MTSummit/papers.html
(consulted 03.09.2007).

- Dabbadie, Marianne et al (2002). "A Hands-On Study of the Reliability and Coherence of Evaluation Metrics". In *Proceedings of LREC 2002 (Language Resources and Evaluation Conference*). Las Palmas, Canary Islands, 27 May 2002. 8-16. Online at: http://mt-archive.info/LREC-2002-Dabbadie-2.pdf (consulted 15.08.2007).

- Doddington, George (2002). "Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics". In *Proceedings of HLT 2002 (2nd Conference on Human Language Technology)*. San Diego, California, 128-132.

- Dyson, Mary and Hannah, Jean (1987). "Toward a Methodology for the Evaluation of Machine Assisted Translation Systems". *Computers and Translation* 2(3), 163-176.

- Hamon, Olivier et al (2007). "Assessing Human and Automated Quality Judgments in the French MT-Evaluation Campaign CESTA". In *Proceedings of*

Machine Translation Summit XI. Copenhagen, Denmark. 10-14 September. 231-238.

- Hutchins, John and Somers, Harold (1992). *An Introduction to Machine Translation*. London: Academic Press Limited.

- King, Margaret, Popescu-Belis, Andrei and Hovy, Eduard (2003). "FEMTI: Creating and Using a Framework for MT Evaluation". In *Proceedings of Machine Translation Summit IX*, 23-27 September, New Orleans. 224-231. Online at: http://www.amtaweb.org/summit/MTSummit/papers.html (consulted 03.09.2007)

- Krings, Hans P. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Geoffrey S. Koby (ed/translator). Kent, Ohio: The Kent State University Press.

- Maegaard, Bente (ed.) (2007). *Proceedings of Machine Translation Summit XI*, 10-14 September 2007. Copenhagen, Denmark.

- Mitamura, Teruko (1999). "Controlled Language for Multilingual Machine Translation". In *Proceedings of Machine Translation Summit VII*. 13-17 September. Kent Ridge Digital Labs, Singapore. 46-52. Online at: www.lti.cs.cmu.edu/Research/Kant/PDF/MTSummit99.pdf (consulted 18.12.2007)

- Mitamura, Teruko and Nyberg, Eric (1995). "Controlled English for Knowledge-Based MT: Experience with the KANT System". In *Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95)*. 5-7 July 1995. Leuven, Belgium, 158-172.

- O'Brien, Sharon (2006). *Machine-Translatability and Post-Editing Effort: An Empirical Study Using Translog and Choice Network Analysis*. Unpublished PhD Dissertation. Dublin City University.

- — (2005). "Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Text Translatability". In *Machine Translation* 19(1), 37-58.

- Roturier, Johann (2006). *An Investigation into the Impact of Controlled English Rules on the Comprehensibility, Usefulness and Acceptability of Machine-Translated Technical Documentation for French and German Users".* Unpublished PhD dissertation. Dublin City University.

- Schäler, Reinhard (1998). "The Problem with Machine Translation". In Bowker, L. et al (eds) *Unity in Diversity: Recent Trends in Translation Studies*. Manchester: St. Jerome, 151-156.

- Senez, Dorothy (1998). "Post-editing Service for Machine Translation Users at the European Commission". In *Proceedings of Translating and the Computer 20*. London: Aslib.

- Underwood, Nancy and Jongejan, Bart (2001). "Translatability Checker: A Tool to Help Decide Whether to Use MT". In Maegaard, B. (ed.) *Proceedings of the MT Summit VII: Machine Translation in the Information Age*. 18-22 September. Santiago de Compostela, Spain, 363-368.

- Way, Andy and Gawronska Barbara (eds) (2007). *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine*

Translation - TMI 2007. 7-9 September. Skövde, Sweden. Online at: http://www.mt-archive.info/TMI-2007-TOC.htm (consulted: 07 January 2008).

- White, John (2003). "How to Evaluate Machine Translation". In Somers, H. (ed.) *Computers and Translation: A Translator's Guide*. Amsterdam and Philadelphia: John Benjamins. 211-244.

Biographies

Dr. Sharon O'Brien is a lecturer in Translation Studies at the School of Applied Language and Intercultural Studies in Dublin City University. She lectures in topics such as translation practice, localisation and translation theory. Her research interests include translation technology, specifically machine translation, controlled language and translation memory tools, translation process and eye-tracking.

E-mail: sharon.obrien@dcu.ie

Rebecca Fiederer has an MA in Translation Studies from Dublin City University. She is currently working as a sub-editor.

Appendix A – Small Sample of the Source Sentences and Translations/Post-Edited Sentences

Source Text	**The user can use the same source files for this operation.**	**ArborText Epic allows the author to specify an explicit structure for documents.**	**From within Windows Explorer, double-click on a document(s) or entity.**	**Create a new document by doing the following.**
Target Text 1	Für diese Operation kann der Benutzer dieselben Ausgangsdateien verwenden.	Mit ArborText Epic kann ein Autor eine explizite Struktur für Dokumente angeben.	Doppelklicken Sie in Windows Explorer auf ein Dokument oder eine Entität.	Erstellen Sie ein neues Dokument. Gehen Sie folgendermaßen vor:
Target Text 2	Der Benutzer kann dieselben Quellendaten für diesen Vorgang verwenden.	ArborText Epic erlaubt dem Autor, eine explizite Struktur für Dokumente anzugeben.	Klicken Sie im Windows Explorer doppelt auf ein oder mehrere Dokument oder eine Entität.	Erstellen Sie ein neues Dokument, indem Sie folgende Schritte ausführen:
Target Text 3	Der Benutzer kann für diesen Arbeitsgang dieselben Quellendateien verwenden.	ArborText Epic erlaubt dem Autor, eine explizite Struktur für Dokumente zu definieren.	Doppelklicken Sie im Windows Explorer auf das gewünschte Dokument oder die gewünschte Entität.	Erstellen Sie ein neues Dokument, indem Sie folgendermaßen vorgehen:
Target Text 4	Hierfür können dieselben Quellendateien verwendet werden.	ArborText Epic gibt dem Benutzer die Möglichkeit, Dokumenten eine explizite	Vom Windows Explorer aus: durch Doppelklicken auf einem oder mehreren Dokumenten	Um ein neues Dokument zu erstellen, gehen Sie wie folgt vor:

		Struktur zu geben.	oder einer Entität.	
Target Text 5	Der Benutzer kann dieselben Quellendaten für diesen Zweck verwenden.	ArborText Epic erlaubt dem Autor, eine explizite Struktur für Dokumente anzugeben.	Von innerhalb des Windows Explorers klicken Sie doppelt auf ein Dokument oder eine Entität.	Erstellen Sie ein neues Dokument, indem Sie Folgendes ausführen.
Target Text 6	Für diese Operation kann der Benutzer die gleichen Quellendateien verwenden.	Arbor Text Epic ermöglicht dem Verfasser, eine explizite Struktur für Dokumente anzugeben.	Klicken Sie im Windows-Explorer doppelt auf ein Dokument oder ein Objekt.	Erstellen Sie ein neues Dokument, indem Sie die folgenden Schritte ausführen:

[1] The imbalance between post-edited and translated sentences arises from the fact that O'Brien's main focus was on logging the effort involved in post-editing, not on a comparison between translation and post-editing.

[2] Translation by R. Fiederer.