Baltic J. Modern Computing, Vol. 4 (2016), No. 2, 141-151

Comparing Translator Acceptability of TM and SMT Outputs

Joss MOORKENS, Andy WAY

ADAPT Centre, School of Computing, Dublin City University, Ireland

joss.moorkens@dcu.ie, away@computing.dcu.ie

Abstract. This paper reports on an initial study that aims to understand whether the acceptability of translation memory (TM) among translators when contrasted with machine translation (MT) unacceptability is based on users' ability to optimise precision in match suggestions. Seven translators were asked to rate whether 60 English-German translated segments were a usable basis for a good target translation. 30 segments were from a domain-appropriate TM without a quality threshold being set, and 30 segments were translated by a general domain statistical MT system. Participants found the MT output more useful on average, with only TM fuzzy matches of over 90% considered more useful. This result suggests that, were the MT community able to provide an accurate quality threshold to users, they would consider MT to be the more useful technology.

Keywords: Machine Translation, Human Evaluation, Translation Memory, Confidence Estimation

1. Introduction

The role of the translator has changed considerably over the past 25-30 years, with technology playing an ever more vital role in a specialised translator's workflow. Bota et al. (2013) noted that some translation technology tools are "more highly regarded than others". Translation Memory (TM), for example, is considered acceptable and necessary (Heyn, 1998), whereas Machine Translation (MT) remains unpopular among many translators. Surveys support the first of these claims, in that while users may have problems with certain extrinsic aspects of TM tools such as pricing or user-friendliness, they have no objection to leveraging previous human translations (Lagoudaki, 2008; Kelly et al., 2012).

For those of us with longer memories, it was not always this way. When commercial TM tools were first introduced, many translators resented the imposition of this new technology. However, early adopters found that, once past the initial learning curve, they could achieve perceptible productivity gains, although the financial benefit of these gains was mitigated to an extent when discounts based on TM matches became common (García, 2006).

As regards the second claim above, a disadvantage for MT is that in exactly the same way as with the introduction of TM, translators further resent the imposition of the newer technology, especially when associated discounts are expected immediately. Translators have complained about having to make tedious repetitive corrections to MT output, lack of creativity, and "limited opportunity to create quality" when post-editing

(Moorkens and O'Brien, 2015). These complaints are exacerbated by the perishable and often poorly-written source content that is pushed towards MT in localisation workflows (Way, 2013; Moorkens and O'Brien, 2015). Despite many studies having shown that MT post-editing increases productivity, users do not always perceive this increase (Koehn, 2009; Gaspari et al., 2014). Despite the increasing incorporation of MT into translation workflows via post-editing (PEMT) or sub-segment auto-suggestion, MT does not yet appear to be widely accepted by translators (cf. Penkale and Way, 2013; Way, 2013).

While this is obvious to many, we consider it worth pointing out the main difference between TM and MT: namely that while MT attempts to translate all sentences in an input document, TM does not (except in the case of 100% matches, for which translators receive little or no remuneration in any case); TM systems merely search the source side of a set of translation pairs for the closest-matching instances above some pre-determined threshold imposed by the translator (so-called 'fuzzy matches'; Sikes (2007)). A ranked list of the said translation pairs is then presented to the translator with user-friendly colour-coding to help the user decide which parts are useful in the composition of the target translation, and which should be ignored and discarded. The addition of project-specific or historical information from the suggested TM segment metadata may help the translator with this decision (Teixeira, 2014). Accordingly, we note the different roles played by the human-in-the-loop here: when using TM, the human still *translates*, whereas with MT, the MT output is usually *post-edited*. There are exceptions here as the delineation between TM and MT has become somewhat blurred, with some tools incorporating both technologies and others adding sub-segment autosuggestions from MT output (Green et al. 2014; O'Brien and Moorkens, 2014).

Given that today's statistical MT (SMT) engines have greatly improved in terms of the quality of their output (cf. Way (2013) for a list of use-cases where MT demonstrably plays an invaluable role), it is disappointing for MT developers to learn that human translators still appear to draw greater satisfaction from slow, interactive TM tools as opposed to fully automatic, fast MT systems. For example, 75% of respondents to Moorkens and O'Brien's (2016) survey of translators agreed that TM helps with their work, whereas only 30% said the same of MT; what's more, 56% indicated that they considered MT a problematic technology. Participants in another study by Moorkens and O'Brien (2015) said that they found post-editing tiring as they are required to be "constantly vigilant … due to the absence of any confidence indication".

More positively, Koskinen and Ruokonen (2016) suggest that translators are "quite willing to adopt new technology as long as it makes their work more efficient". Consequently, we feel that some of the problems with MT reside in how it is presented. In particular, if making productivity improvements could be made demonstrable to and perceptible for users, there would be far fewer objections to MT as a technology in its own right than we have seen heretofore. Accordingly, this paper reports on an initial study that seeks to answer the following question: *Is comparative acceptability of TM over MT predicated on the user's ability to optimise the precision and usefulness of match suggestions by setting a minimum match threshold?*

We contend that the answer to this question is yes, and that:

1. MT would be considered more acceptable to users if only those matches that required relatively small amounts of editing were presented to post-editors.
2. TM would be less acceptable to users if matches that required large amounts of editing were presented to translators.

In other words, we suggest that translators' comparative preference of TM over MT demonstrates their preference for precision over maximum recall.

Guerberof (2012) found that the average post-editing time for English-to-Spanish MT trained on a source text-appropriate technical domain was roughly equivalent to the time required to edit an 85-94% fuzzy TM match. Not all MT output will be of this quality of course, and the translator's bugbear of repetitive mistakes to correct in MT output remains a problem, although great strides are being made to incorporate translator feedback into iterative retraining of SMT systems (cf. Du et al., 2015). However, if the ability to set an accurate threshold is one of the things that makes TM useful and acceptable to translators, this highlights the need for accurate confidence prediction of MT quality that correlates with human judgement (e.g. Specia et al. (2009); Specia (2011); Turchi et al. (2013)), the absence of which we believe to be a major stumbling block for acceptability.

This study is reported with the caveat that the research was carried out with a small number of translators for a single domain and language pair, and is intended to preface a larger-scale study that will include measures of actual post-editing effort. However, all participants have substantial translation experience (on average 11.4 years of professional translation experience and 4.5 years of professional post-editing experience) and the chosen language pair of English-German is acknowledged to be a difficult one for MT systems.

The remainder of this paper is organised as follows. In Section 2, we describe the methodology chosen to test the central hypothesis in this paper. In Section 3, we present the results of the experiments conducted, which are discussed further in Section 4. In Section 5, we conclude, and list a number of avenues for further work in this area.

2. Methodology

In this study, seven translators were asked to rate the usefulness of 60 match suggestions in German for 60 English source text segments. Source segments were taken from the documentation for the open-source computer-aided design (CAD) program FreeCAD and from the Wikipedia page for CAD.[1] Table 1 shows the homogeneity of all segments, the segments used for TM matching, and those translated using MT, which all exhibited similar characteristics (well within the standard deviation for each text) using common corpora analyses (such as the type/token ratio of lexical variation) in the WordSmith WordList tool.[2] Note that some types appear in both TM and MT corpora.

Table 1. Wordsmith statistics for source data.

	Overall	**Segments for MT**	**Segments for TM**
Types (distinct words)	447	260	268
Type/token ratio (TTR)	42.21	46.93	53.07
Mean word length (chars.)	4.82	4.87	4.76

[1] https://en.wikipedia.org/wiki/Computer-aided_design
[2] http://www.lexically.net/wordsmith

30 segments were translated into German using the generic Microsoft Bing SMT system[3] and 30 target segments were fuzzy match suggestions offered by the Omega-T[4] tool loaded with an English-German TM. The TM was created from the translation of documentation from a commercial CAD software tool, and contained 301,583 translation units (although 7,659 of these contained only numbers, dates, or punctuation symbols). The TM tool suggested only a few matches – there were 42 matches for 141 segments, a match rate of just 29.8% – with most of those suggestions having a low fuzzy match score.[5] For this reason, matches were not taken sequentially, but chosen to provide a reasonable variety of fuzzy match scores. The top-10 fuzzy matches ranged from 73 to 100% and the lowest 10 from 19 to 46%. The range of fuzzy matches are shown in Figure 1.

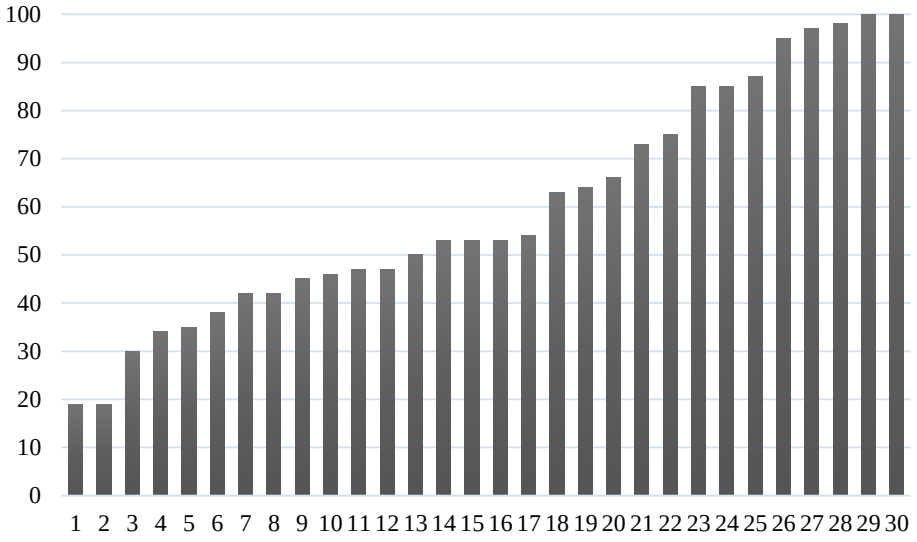

Figure 1. TM target segments' fuzzy-match percentage.

Source text segments and their associated TM or MT target segments were randomised and copied into a six-page survey,[6] where each page contained 10 target text suggestions without any indication of provenance or quality. Participants – all of whom were paid – were informed about the background of the study, and that they could withdraw at any time without penalty (although none did). They were then asked to fill in details of their translation experience, age range, and opinion of MT (all non-mandatory questions) before beginning to rate the 60 segments. Ratings were based on a decision as to whether to retain or delete the target suggestion before beginning to edit or translate from scratch, and were similar to those used by Krings (2001) and Specia et

[3] http://www.bing.com/translator/
[4] http://www.omegat.org/
[5] We take this as supporting evidence of our claim that TM technology is actually of little use to most translators, and certainly nowhere near as potentially useful as MT.
[6] The survey used the Limeservice platform, available at www.limeservice.com.

al. (2009), and modified from the rating descriptions used in Moorkens et al. (2015), which were found to be an inconsistent predictor of post-editing effort. The ratings chosen by participants via radio button for each segment were as follows:

1. Not usable – delete and translate from scratch,
2. Useful – editing is faster than translation from scratch,
3. Almost perfect – only requires minor edits or none at all.

This study used purposive sampling, gathering participants appropriate to the research question. Participants were requested to take part via an open call on social media and direct emails to translators with the appropriate language pair listed on the website of the Irish Translators and Interpreters Association.[7] Participants reported between one and 22 years' translation experience. Five participants had experience of post-editing of between five months and 11 years. Users' attitudes to MT tended to be positive, with one considering it "useful for repetitive texts". Users did not consider MT a threat to translators, an attitude consistent with translators in other studies (Katan, 2011). One participant suggested that they are complementary technologies and said "I highly doubt MT can ever replace HT". Another wrote: "Some think MT will replace [human] translation, but although it's getting better and better, that is not possible for the majority of content out there".[8] Another participant said that MT is an "excellent productivity tool when used for suitable content", and that its "greatest advantage is the often higher consistency in terminology and style". Two participants were less effusive, with one writing that it's only useful for "technical texts with a simple sentence structure", and another considering that he or she works faster without MT, which is "not usable for professional translations without heavy editing". We have to acknowledge the possibility that translators with a very negative opinion of MT chose not to take part on the basis of the project description in the emailed invitation to participate, although we do not consider this to be very likely.

3. Results

Participants spent on average 31.6 minutes (max. 53 minutes 14 seconds, min. 14 minutes 28 seconds) completing the rating survey, including one participant who completed the survey in two sittings on consecutive days. Inter-rater agreement was considered moderate using Fleiss' kappa, where K=0.446. Percentage agreement was 77.8% overall, with greater consistency amongst raters for TM matches (84.3%) than for MT output (71.3%), a more consistent result than in Moorkens et al. (2015), albeit with fewer participants.

On average, participants rated the MT output more positively as a basis for post-editing, with a median segment rating of 2 (where 1 is not usable and 3 is almost

[7] http://www.translatorsassociation.ie/

[8] It is refreshing to see well-informed translators speaking with authority on MT, owing the great strides taken by the MT community to reach out to translators on this issue. It is all the more disappointing, then, to see TAUS' recent blog "The Future Does Not Need Translators" (https://www.taus.net/blog/the-future-does-not-need-translators) which in our opinion seeks to undermine the *status quo* and unnecessarily antagonise translators.

perfect). It is interesting to note that no machine-translated segment received a rating of 1 from all participants. The median rating for TM matches without a fuzzy match threshold was 1.14. Segments were randomised and presented without any indication of fuzzy match percentage, or whether the target text came from MT or TM. Despite this, there was a very strong correlation between fuzzy match percentage and average participant rating, where r=0.838 (and $p < 0.001$). Table 2 shows how many times each rating was chosen by a participant for segments from MT and TM.

Table 2. Number of occurrences of each rating.

Rating	Overall	TM	MT
1	185	136	49
2	146	37	109
3	89	37	52

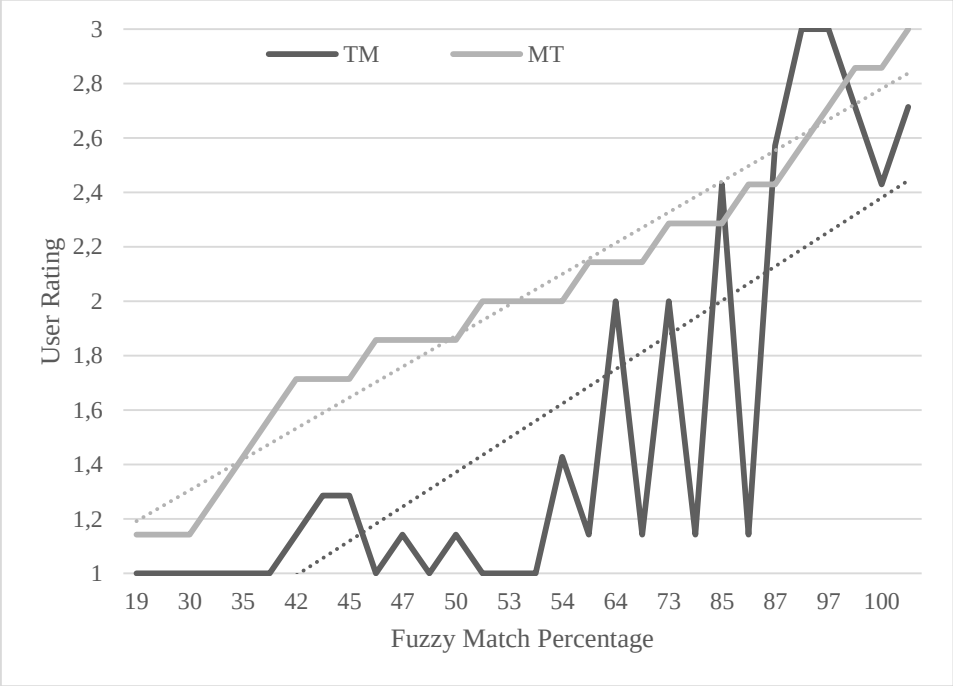

Figure 2. Relationship between average rating and fuzzy match percentage for TM. Ratings for MT output are shown for comparison, charted in order of average rating.

Figure 2 shows the relationship between fuzzy match percentage and average participant rating. Ratings for MT output are shown for comparison, charted in order of

average rating (as they had no fuzzy match percentage). Trend lines for TM and MT output clearly show the comparatively higher quality of MT output amongst participants. Note too that at 70-75% fuzzy match thresholds, settings which are often applied in practice (SDL Trados Studio's default threshold is 70%), many TM matches are ranked well below equivalent MT suggestions, with many ranked as not useable at all. This further demonstrates (cf. Simard and Isabelle, 2009) that arbitrarily imposing a cut-off as is the norm in the translation industry above which TM is used and below which MT is used, harms translator performance.

Participants were presented with a free text box after every ten segments in which to make comments. One participant mentioned the amount of time required to make a quality judgement on a proposed target text segment. He wrote that "MT Output has to make some kind of general sense (syntactically) to trigger within the post-editor a positive impulse to start post-editing and not dismiss [it]". Another participant commented that "the main thing to keep in mind is that the analysis of MT output takes time and the smart decision is to dismiss such segments early." De Almeida (2013) believes that this decision-making at speed is problematic for some translators, especially for mid-ranking MT outputs. Koehn (2009), when discussing periods spent by translators pausing during the translation process, notes that "different lengths of pauses indicate the different problems which the translators are dealing with", and that for post-editors "most of the time is spent on contemplating changes, but very little on executing them". Speculating about the translator's behaviour during such pauses, Koehn (2009) intuits that the translator "is reading more of the MT output and looking for mistakes to be corrected". Mesa-Lao (2014) also stresses this focus on the target text for post-editing, noting that study participants either give the source text a cursory read or skip "straight to the target text in search of errors". In this study we found a moderate correlation between the number of mid-ranking (ranked 2: "useful – editing is faster than translation from scratch") segments on a survey page of ten and the amount of time (in comparison to the participant's median time) required for completing a survey page of ten ranking exercises (r_s=0.44141, p=0.006). We suggest that while some translators can decide quite quickly whether to accept and post-edit good MT outputs, and reject poor MT output in favour of translating themselves from scratch, those of a middling quality slow down the translators' decision-making process, exactly what we are trying to avoid by introducing technology into the translation pipeline.

11 of the 30 TM match proposals in this study received scores of 1 (not usable) from all participants, suggesting that they considered the matches wholly dissimilar to the source text, despite five of these proposals receiving match percentages of around 50%. One participant betrayed some irritation with these poor TM proposals, commenting that they contained "serious mistranslations that cannot be understood without reading the source a few times." In a small number of instances, TM match proposals were dismissed when they might have been used as a basis for editing, or might have been perceived more favourably when displayed with a high fuzzy match percentage or with sections for leverage highlighted or colour-coded within a TM tool. For example, the proposed target text for the source segment '*Cmd-1 turns the Tool Sets palette on and off.*' was '*Aktiviert und deaktiviert den Fangmodus'* [Activates and deactivates the snap mode]. Here the 'activates and deactivates' phrase could have been leveraged by users, but all participants considered the segment as a whole unusable and eschewed this option.

4. Discussion

The strong association between fuzzy match percentages and participant ratings, despite the fact that percentages were not displayed onscreen, demonstrates an advantage that TM has over MT: fuzzy matches are reasonably accurate gauges of quality that correlate with human judgement, whereas "the correlation between human judges and all [contemporary] automatic measures of MT quality" is "quite low" (Turian et al., 2003). Some progress has been made in research on MT confidence estimation without use of reference translations. Specia et al. (2009), in a study that aimed to eliminate very poor MT results, identified 84 segment-level features that could be used to estimate MT quality, with results that correlated far better with human judgements than several commonly-used automatic evaluation metrics. Accurately gauging the quality of time-consuming mid-ranking MT output will be a more onerous task. Turchi et al. (2013) noted the subjectivity of human judgements and the associated difficulty in confidence estimation using machine learning based on human annotation. Specia (2011) suggested machine-learning models based on user post-edits as a route for accurate MT confidence estimation. It is less likely that users would have accepted TM were it not possible to impose a quality threshold that users can confidently consider accurate and personalise to their own requirements based on years of experience. Once this threshold is removed, participants in this study commented on the low-quality match proposals and rated many segments poorly. For this reason, we consider the answer to the research question presented in Section 1 to be answered – at least in part – in the affirmative, such that comparative acceptability of TM over MT is indeed predicated on the user's ability to set a minimum fuzzy match threshold.

In Section 1, we mentioned users' complaints that MT output requires "constant vigilance", but results in Section 3 also highlighted the time required for manual evaluation of MT output prior to post-editing. TM tools not only provide users with accurate measures of quality, indicating words and colour-coded sub-segments that may be left untouched, but the time and effort required for manual evaluation is also removed. This suggests that the ability to set an accurate threshold for MT quality (which would be made easier if an automatic metric can be found that correlates strongly with human judgement) should lead to further productivity gains by saving the time required for manual evaluation, as well as reducing the user's cognitive effort.

Note that this is harder than it might seem, as translators are not necessarily good arbiters of MT quality, especially vis-à-vis TM quality. In their work on combining SMT and TM for optimal translation recommendation to post-editors, He et al. (2010) note in their evaluation that while end-users are not made aware of which segments come from SMT and which from TM, one post-editor "obviously mistakes MT outputs for TM outputs". They note that this indicates not only that "phrase-based SMT system[s] [are] able to produce outputs that are … grammatically acceptable enough to be recognized as human translations in the TM", but also "how much the post-editors subconsciously trust the TM [which] may be an explanation for the relatively low acceptance of MT technology in the localization industry and demonstrates the need for TM–MT integration". In a similar line of work, we note here the recent effort by STAR to combine TM and MT in an interesting way, where MT matches are used to reinforce fuzzy matching (Hofmann, 2015).

Participants in this study were reasonably satisfied with the quality of MT output despite the use of a generic engine for a difficult language pair. This suggests that,

contrary to perceived wisdom in the field, quality is not the sole barrier for widespread MT acceptance. The results of Moorkens et al. (2015) showed that the addition of onscreen MT confidence indication alone does not immediately lead to behavioural changes for post-editors. Users need to learn to trust measures of quality or confidence, but also need to be presented only with proposals (as segments or sub-segments) that will be useful to them.

Participants in this study were mostly well-disposed to MT, and as such were willing to rate segments without prejudice, despite the lack of provenance metadata. They had the confidence to participate in MT research without being suspicious of the research motives. The challenge will be to convince those less well-disposed to MT that automatic translation can be perceptibly beneficial, despite the increasingly large body of evidence to support this point of view. As a move in that direction, we suggest that the ability to only display useful MT output will greatly improve acceptability.

5. Conclusions and Further Work

In this paper, we have set out to challenge the perception among translators that TM is a more useful technology than MT is. While this does not appear to be true *per se*, what is unquestionably important is the translator's ability to control the fuzzy match threshold. When low- and mid-ranking fuzzy matches are presented to translators without the accompanying fuzzy match scores, translators find the suggestions irritating, and for over 36% of such instances, useless for their purposes. In contrast, *all* of the MT matches suggested were rated as having some utility to post-editors.

Accordingly, we contend that this finding demonstrates very clearly a serious mistake that has been made by introducing MT into the PEMT pipeline. Translators are quite used – one might even say 'happy' – to not having help from CAT tools for every segment, as TM offers useful suggestions only some of the time; in this study we found fuzzy matches for 29.8% of segments, although many of these at 13-70% (see Figure 1) could not be considered useful. In contrast, MT developers have allowed the soft underbellies of their engines to be exposed 'warts and all' to translators, as MT outputs are typically provided for every source segment. What we have demonstrated in this paper is that when the constraints on fuzzy match thresholding are relaxed, translators actually find TM to be of (much) less use than SMT. This suggests to us very strongly that robust, reliable MT confidence measures need to be developed as a matter of urgency which can be used by post-editors to wrest control over what MT outputs they wish to see, and perhaps more importantly still, which ones should be withheld.

In further work, we aim to extend this study to more language pairs, and larger amounts of data translated using *both* TM and MT, assigned randomly to a wider range of translators not only for rating, but also for post-editing. We expect the conclusions drawn in this initial study to be confirmed in further research, which will, we hope, concentrate the minds of MT engine developers to develop a consistent measure of quality for each MT segment output by the system, which can be relied upon – and configured – by human translators. We expect that once translators can control what MT output they actually get to see, then MT will meet with considerably wider acceptance from the translator community.

Acknowledgements

This research is supported by the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded under the European Regional Development Fund. The authors wish to extend their thanks to the participants in this research, and to the anonymous reviewers whose suggestions and advice were gratefully received.

References

Bota, L., Schneider, C., and Way, A. (2013). COACH: Designing a new CAT Tool with Translator Interaction. *Proceedings of Machine Translation Summit XIV,* (Nice, France), 8pp.

De Almeida, G. (2013). *Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two Romance languages.* PhD Thesis. Dublin City University, Ireland.

Du, J., Srivastava, A., Way, A., Maldonado-Guerra, A., and Lewis, D. (2015). An Empirical Study of Segment Prioritization for Incrementally Retrained Post-Editing-Based SMT. *MT Summit XV, Proceedings of the Fifteenth Machine Translation Summit*, (Miami, FL), 172—185.

García, I. (2006). Translation Memories: A Blessing or a Curse? In Pym, A., Perekrestenko, S. (Eds.) *Translation Technology and its Teaching*. Tarragona, Spain: Universitat Rovira i Virgili.

Gaspari, F., Toral, A., Kumar Naskar, S., Groves, D., and Way, A. (2014). Perception vs Reality: Measuring Machine Translation Post-Editing Productivity. *Proceedings of the 11th conference of the Association for Machine Translation in the Americas: Workshop on Post-editing Technology and Practice (WPTP3)*, (Vancouver, Canada), 60–72.

Green, S., Wang, S., Chuang, J., Heer, J., Schuster, S., and Manning, C. D. (2014). Human Effort and Machine Learnability in Computer Aided Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), 1225—1236.

Guerberof, A. (2012). *Productivity and Quality in the Post-editing of Outputs from Translation Memories and Machine Translation*. PhD thesis. Universitat Rovira i Virgili, Tarragona, Spain.

He, Y., Ma, Y., Roturier, J., Way, A., and van Genabith, J. (2010). Improving the Post-Editing Experience Using Translation Recommendation: A User Study. *Proceedings of AMTA 2010: The Ninth Conference of the Association for Machine Translation in the Americas*, (Denver, CO.), 247—256.

Heyn, M. (1998). Translation Memories – Insights & Prospects. In L. Bowker, M. Cronin, D. Kenny and J. Pearson (Eds.) *Unity in Diversity? Current Trends in Translation Studies*, Manchester: St. Jerome, 123—136.

Hofmann, N. (2015). MT-enhanced fuzzy matching with Transit NXT and STAR Moses. *EAMT-2015: Proceedings of the Eighteenth Annual Conference of the European Association for Machine Translation*, (Antalya, Turkey), 215.

Katan, D. (2011). Occupation or profession: A survey of the translators' world. In R. Sela-Sheffy & M. Shlesinger (Eds.), *Profession, identity and status: Translators and interpreters as an occupational group*, Amsterdam: John Benjamins, 65-88.

Kelly, N., DePalma, D. A., and Hegde, V. (2012). Voices from the freelance translator community (Report). *Common Sense Advisory*, Boston MA.

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

Koskinen, K., and Ruokonen, M. (2016). Love letters or hate mail? Translators' technology acceptance in the light of their emotional narratives. In D. Kenny (Ed.), *Human Issues in Translation Technology: The IATIS Yearbook*. Abingdon: Routledge (to appear).

Krings, H. P. (2001). *Repairing Texts*. Kent State University Press, Ohio, USA.

Lagoudaki, E. (2008). *Expanding the possibilities of translation memory systems*. PhD thesis. Imperial College, London, UK.

Mesa-Lao, B. (2014). Gaze behaviour on source texts: An exploratory study comparing translation and post-editing. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, Lucia Specia (Eds.), *Post-editing of Machine Translation: Processes and Applications*, Newcastle-upon-Tyne: Cambridge Scholars, 219-245.

Moorkens, J., and O'Brien, S. (2015). Post-Editing Evaluations: Trade-offs between Novice and Professional Participants. *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT 2015),* (Antalya, Turkey), 75—81.

Moorkens, J., O'Brien, S., Silva, I. A. L., Fonseca, N., and Alves, F. (2015). Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation* 29(3-4), 267-284. doi: 10.1007/s10590-015-9175-2

Moorkens, J., and O'Brien, S. (2016). Assessing User Interface Needs of Post-Editors of Machine Translation. In Dorothy Kenny (Ed.*), Human Issues in Translation Technology: The IATIS Yearbook.* Abingdon: Routledge (to appear).

O'Brien, S., and Moorkens, J. (2014). Towards intelligent post-editing interfaces. In Baur, W., Eichner, B., Kalina, S., Kessler, N., Mayer, F. and Orsted, J. (Eds.) *Man versus Machine: Proceedings of the XXth FIT World Congress (Vol. I)*, Berlin, Germany: BDÜ, 131—137.

Penkale, S., and Way, A. (2013). Tailor-made Quality-controlled Translation. *Proceedings of Translating and the Computer 35*, London, UK, 7pp.

Sikes, R. (2007). Fuzzy matching in theory and practice. *Multilingual*, 18(6):39 – 43.

Simard, M. and Isabelle, P. (2009). Phrase-based machine translation in a computer-assisted translation environment. *Proceedings of The Twelfth Machine Translation Summit (MT Summit XII),* (Ottawa, Canada), 120 – 127.

Specia, L., Cancedda, N., Dymetman, M., Turchi M., and Cristianini, N. (2009). Estimating the Sentence-Level Quality of Machine Translation Systems. *Proceedings of the 13th Annual Conference of the EAMT*, (Barcelona, Spain), 28–35.

Specia, L. (2011). Exploiting objective annotations for measuring translation post-editing effort. *Proceedings of the 15th conference of EAMT*, (Leuven, Belgium), 73–80.

Teixeira, C. S. C. (2014). The handling of translation metadata in translation tools. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, Lucia Specia (Eds.), *Post-editing of Machine Translation: Processes and Applications*, Newcastle-upon-Tyne: Cambridge Scholars, 109—125.

Turchi, M., Negri, M., and Federico, M. (2013). Coping with the Subjectivity of Human Judgements in MT Quality Estimation. *Proceedings of the 8th Workshop on Statistical Machine Translation (WMT'13),* (Sofia, Bulgaria), 240—251.

Turian, J. P., Shen, L., and Melamed, I. D. (2003). Evaluation of Machine Translation and its Evaluation. *Proceedings of MT Summit IX*, (New Orleans, U.S.A), 386-393.

Way, A. (2013). Traditional and Emerging Use-Cases for Machine Translation. *Proceedings of Translating and the Computer 35*, London, UK, 12pp.

Received May 2, 2016, accepted May 8, 2016