Translators' Perceptions of Literary Post-editing using Statistical and Neural Machine Translation

(Article to be published in Translation Spaces issue 7:2, November 2018)

Joss Moorkens | Dublin City University
Antonio Toral | University of Groningen
Sheila Castilho | Dublin City University
Andy Way | Dublin City University

Abstract
In the context of recent improvements in the quality of machine translation (MT) output and new use cases being found for that output, this article reports on an experiment using statistical and neural MT systems to translate literature. Six professional translators with experience of literary translation produced English-to-Catalan translations under three conditions: translation from scratch, neural MT post-editing, and statistical MT post-editing. They provided feedback before and after the translation via questionnaires and interviews. While all participants prefer to translate from scratch, mostly due to the freedom to be creative without the constraints of segment-level segmentation, those with less experience find the MT suggestions useful.

Keywords
Machine translation, literary machine translation, neural machine translation, human factors in machine translation

Introduction

Machine translation (MT) is being increasingly used by translators as a productivity tool. Although translators have post-edited MT for many years, early use-cases focused on very narrow domains such French-English meteorological texts (Thouin 1982) and English-Spanish public health texts (Vasconcellos 1985), whereas more recently post-editing (PE) has gone mainstream (Lommel and DePalma 2016) with more use-cases being found for post-edited MT (Way 2013). The orthodoxy for the use of MT and PE has been that its incorporation should be in line with the perishability of the text to be translated (Way 2018a). MT is considered to work best on short, unambiguous source text sentences that require a translation that is literal (i. e. translation that is word-for-word or adheres closely to the source text) rather than creative. Accordingly, MT is commonly used without further human intervention for digital texts with a short lifespan, such as online documentation, reviews, and blog posts.

A further limitation to the broader use of MT is that PE is a task disliked by many translators, who have complained that it constrains their work, allows limited opportunities for creativity, and forces them to correct 'stupid' errors repeatedly (Cadwell, Castilho, O'Brien, and Mitchell 2016, Moorkens and O'Brien 2017). The rule-based and data-driven MT systems primarily used until recently are also known to produce literal translations (Martín and Serra 2014). PE has been found to 'prime' the translator, resulting in a final translation that is likely to be similar to the MT

suggestion (Green, Heer, and Manning 2013). Despite this, many studies have found PE to be a useful tool to increase productivity while maintaining acceptable translation quality (Guerberof 2012, Daems, De Clercq and Macken 2017).

Recent years have seen the advent of Neural Machine Translation (NMT), a statistical paradigm for translation using neural networks (Forcada 2017). As well as producing more fluent translations that contain fewer errors (Klubička, Toral, and Sánchez-Cartagena 2017, Way 2018b), NMT has been found to produce less literal translations than the previously dominant paradigms (Castilho et al. 2017a). These findings are based on evaluations to date, which have been carried out using technical and educational texts. The current study has two objectives: firstly, to assess translators' attitude towards PE of NMT output, about which little or no research has yet been published, and secondly, to test the capability of a state-of-the-art NMT system to aid PE of literary texts, a challenging long-lifespan text type, to which MT and PE is not usually applied. We compare levels of PE effort and perceived post-task acceptability when translating from scratch, post-editing phrase-based Statistical MT (SMT), and post-editing NMT. Our hypothesis is that NMT PE will be more productive and acceptable than SMT PE, but that participants will prefer to translate from scratch, as literary PE is not yet a common task.

We consider an evaluation of literary MT and PE timely due to the advent of NMT, for which claims have been made regarding high quality and the ability to place translated words in the appropriate context, and the growing availability of e-books. These e-books are an ideal resource for training literary-adapted MT systems, both using monolingual data (novels in a digital format) and bilingual or parallel data (digital novels and their translations). Literary translation will test the ability of NMT systems to efficiently produce translations that, followed by PE, move beyond the limited quality expectations of perishable texts.

To achieve our objectives, we set six professional translators with experience in literary translation the task of post-editing a chapter of a literary novel that has been automatically translated from English to Catalan, measuring the level of PE effort (reported in detail elsewhere in Toral, Wieling, and Way 2018). Each participant translator was asked to translate sections of the novel under three different conditions: translation from scratch (the usual process for literary translation), SMT PE, and NMT PE. This article focuses on the translators' perceptions of the task, whether they found the MT suggestions useful, whether they would consider post-editing literary texts in the future, and contrasts these task perceptions with the technical and temporal measurements

The following section reviews previous studies of translators' attitudes to MT and PE, and looks at published assessments of NMT quality and previous studies of MT applied to literary texts. Thereafter, the methodology for this research is described in detail, with profiles of participants and details of the MT systems used and the data used to train them. The results sections provide measurements of technical and temporal PE effort, and participants' perceptions of the task. Finally, a discussion section considers the implications of this study and suggests opportunities for further research.

Related literature

Perceptions of post-editing

The practice of post-editing is defined by Wagner (1985: 1) as "correction of a pre-translated text rather than translation 'from scratch'", and more specifically by Somers (2001: 138) as a process of "tidying up the raw output, correcting mistakes, revisiting entire, or, in the worst case, retranslating entire sections" of MT output. PE has received a great deal of research attention in recent years. MT is now integrated with many translation tools (Castilho and O'Brien 2016) and there is growing industrial use due to improving MT quality (e.g. Green, Heer, and Manning 2013), time-focused production cycles, and a need for high throughput despite economic constraints (Moorkens 2017). Wagner (1985: 2) noted that translators may not welcome "working by correction rather than creation", and studies have repeatedly shown PE to be an unpopular task, despite it being associated with productivity gains in almost all published studies (Teixeira 2014, Moorkens and O'Brien 2017). All of the studies in this section are based on PE prior to the shift to NMT – a limitation in this review that serves to highlight the novelty of this article at the time of publication.

Several works have studied the effort involved in PE tasks (De Almeida & O'Brien 2010, Specia 2011, Guerberof 2012, Lacruz and Shreve 2014, Viera 2014, Carl, Gutermuth and Hansen-Schirra 2015, Koponen 2016, Daems et al. 2017). These studies invariably use the categories of PE effort suggested by Krings in 2001: temporal effort (time spent post-editing), technical effort (number of edits, often measured using keystroke analysis or approximated using the TER (Snover et al. 2006) metric), and cognitive effort (often measured using the proxy of fixation data from an eye-tracker or approximated using a think-aloud protocol or pause analysis).

Studies have shown that translators often perceive that they have been less productive when questioned after post-editing, even when they were actually faster some or all of the time (Plitt and Masselot 2010; Gaspari et al. 2014). Teixeira (2014) found that, despite productivity gains, many translators still prefer to translate from scratch. Research on interaction with MT, such as Moorkens et al. (2015), help to explain this by shedding light on the issue of perception of *cognitive* effort by translators. The authors focused on testing the correlation of actual temporal, technical and cognitive PE effort with ratings of perceived effort by the post-editors. They found that the human predictions of PE effort did not correlate well with the actual time required for PE. Measures of cognitive effort, however, did correlate with temporal and technical PE effort.

Human translators' perceptions of translation technology were also studied by LeBlanc (2013:7-9), where the author gathered opinions regarding the advantages and disadvantages of translation memories (TM). On the one hand, translators found that TMs can help to increase productivity and improve consistency (terminology, phraseology), on the other hand they found TMs to be a "barrier to creativity" and that they can "make translators lazy and increasingly passive". In addition, translators found that TMs could "lead to a de-skilling of the translator and thus have an effect on the translator's professional satisfaction and, ultimately, his/her status".

Church and Hovy (1993: 247) called PE an “extremely boring, tedious, and unrewarding chore.” Moorkens and O’Brien (2017: 109) explained further that translators find PE to be an “edit-intensive, mechanical task that requires correction of basic linguistic errors over and over again”. In their study, only 18% of respondents profess to like working with MT. Other authors such as Kelly (2014) have described PE even more negatively as “linguistic janitorial work”.

In other work that studied translators’ perceptions of MT and PE, Cadwell, Castilho, O’Brien, and Mitchell (2016) interviewed translators at the Directorate General for Translation of the European Commission (DGT) to find reasons for the adoption or non-adoption of MT in their work flow. The authors found that the most common reasons to use MT in their everyday work were due to i) productive gains, ii) perceived good quality of the MT output, and iii) for inspiration, new ideas or as a kick-start the translation process. Moreover, translators’ reasons also included “to get a gist understanding” of the source, “to miss fewer elements of source content”, and “for texts which can be easily processed by a machine” (ibid. 236). In contrast, translators’ reasons for not adopting the use of MT include i) perceived poor quality of the MT output or a bad first experience, ii) texts that do not have great retrieval from TMs, such as speeches and press releases, and iii) because of “MT’s negative influence on a translator's abilities”. Other agreed reasons also include the fear of the unknown or “being replaced by a machine”; the fact that translators found that “MT induces the translator to make particular errors” as well as MT errors requiring “extra attention“, but mostly interestingly because using “MT devalues a translator's work”, diminishes creativity and “cannot be trusted”. In follow-up work, Cadwell, O’Brien, and Teixeira (2017:10-11) compare the perceptions of the DGT translators against the perceptions of translators working for a private company. The authors found that the reasons to adopt or not the use of MT in the translators’ daily work flow mostly overlaps to those of the DGT translators, but a few other reasons to adopt MT were highlighted, such as “greater MT adoption is inevitable” and “PE can be creative”; in contrast, other reasons to not adopt were added to the list is, such as “PE is slower than other methods”, “PE is not an enjoyable task” and “PE work is not compensated fairly”.

Neural MT

Neural MT has quite quickly become the dominant paradigm for MT research and large-scale deployment on the basis of strong evaluation results using automatic metrics (Bahdanau et al. 2014, Sennrich et al. 2016a), impressive results in competitive shared tasks (Bojar et al. 2016), and some high-profile research papers (such as Wu et al. 2016, Hassan et al. 2018). Evaluations using human and automatic evaluation, often in comparison with SMT (such as Bentivogli et al. 2017, Castilho et al. 2017a), find NMT output to be more fluent than SMT output, with fewer word-order errors, particularly with regard to verb placement. The most prevalent error types in NMT output appear to be mistranslations and omissions. In general human rating tasks, NMT segments tend to fare better than SMT, but when rated for adequacy (the extent to which the translated text reflects the meaning of the source text) and fluency, adequacy gains are less marked for NMT.

Results for PE effort have also been mixed. Bentivogli et al. (2017; transcribed speeches for English-to-German), Toral, Wieling, and Way (2018; a novel for English-to-Catalan) and Castilho et al. (2017; teaching and learning texts for English-to-Greek, -German, -Portuguese, and -Russian) report that technical effort was reduced when using NMT, but the latter study finds no real improvement in terms of temporal effort when compared with SMT. Toral and Sánchez-Cartagena (2017) find a link between sentence length and NMT quality, reporting that SMT performs better than NMT for segments longer than 40 words. Koehn and Knowles (2017) reiterate this along with other currently unsolved difficulties for NMT development, including problems with out-of-domain segments and low-frequency words. The mostly positive assessments of NMT quality and the fast pace of NMT research and development suggests that this paradigm is now very much the state of the art where sufficient in-domain training data, expertise, and computing power is available.

MT for literature

The establishment of the Computational Linguistics for Literature workshop in 2012 signalled a growing interested in literature among the computational linguistics research community. This has largely related to the automatic identification of text snippets that convey figurative devices. Research on the application of MT to literature has been more limited. Some of this has applied MT to poetry (Genzel et al. 2010, Greene et al. 2010) or poetry with sections of prose (Jones and Irvine 2013). Besacier (2014) suggested that MT and PE (by non-professional translators) might be a useful low-cost alternative to human translation of literary works for those willing to sacrifice a degree of quality.

More recent work by Toral and Way (2015) involved a comparative analysis of the translatability of literary text for a language pair of closely-related languages (Spanish-to-Catalan) according to narrowness of the domain and freedom of translation, as well as a human evaluation of SMT for literary text for that language pair; MT outputs were found to be of equivalent quality to professional human translation for 60% of the segments. Toral and Way (2018) is the forerunner to the current study, in that it uses literary-adapted English-to-Catalan SMT and NMT systems to produce a translation of 12 novels. Human evaluators (native speakers of Catalan) found that NMT outperformed SMT, and with two of the books (*The Catcher in the Rye* and *Harry Potter: The Deathly Hallows*), they perceived NMT translations to be of equivalent quality to human translation for roughly one third of the segments reviewed. The current study is the first to employ PE, for which effort measurements are reported in detail in Toral, Wieling, and Way (2018), and to invite feedback from literary translators.

Methodology

MT systems and training

Two bespoke MT systems, domain-adapted for novels, were trained for this study: a phrase-based SMT system and an NMT system (full details of system training and data pre-processing

can be found in Toral and Way 2018). The in-domain parallel data used for both systems were 133 parallel novels, equating to over 1 million sentence pairs. In-domain monolingual data were roughly 1,000 books written in Catalan, equating to over 5 million sentences.

The SMT system was trained on a combination of in-domain and out-of-domain (around 400,000 sentence pairs of subtitles from Open Subtitles) parallel data, using version 3 of the Moses toolkit (Koehn et al. 2007). The language model for this system is *n*-gram-based, and uses monolingual data, both in-domain and out-of-domain (circa 16 million Catalan sentences crawled from the web). The system uses lexical-, phrase-based, and hierarchical reordering models, along with an operation sequence model (Durrani et al. 2011) and an additional language model based on continuous space *n*-grams (Vaswani et al. 2013), both trained on the in-domain parallel data.

The NMT system, which follows the encoder-decoder approach with attention, was built with Nematus (Sennrich et al. 2017), and uses the same in-domain parallel training data. Out-of-domain data was not used for this system as, at the time of building the system, there was no established method of NMT domain adaptation. NMT systems at the time could not be trained using monolingual data, so instead, a synthetic parallel corpus was created by back-translating the Catalan in-domain monolingual training data (1,000 books) into English (Sennrich, Haddow, and Birch 2016a).[1] The system is trained on sub-word units (a process sometimes called Byte Pair Encoding, proposed for NMT by Sennrich, Haddow, and Birch (2016b) to improve coverage of out-of-vocabulary compound words), whereby the training data are segmented into characters and 90,000 operations were performed between the source and target languages. Finally, an N-best list was generated with the NMT system and reranked with a left-to-right NMT system.

Test data: Warbreaker

The novel used as test data in this study had to be a challenging literary text, and needed to be freely redistributable in order to guarantee the reproducibility of our experiment. The novel chosen was *Warbreaker*, a fantasy novel written by US author Brandon Sanderson.
The book was published under a Creative Commons License (CC-NC-ND specifically), and positive critical reviews attest to a good level of literary quality. The novel (as the training data) was segmented using the Natural Language Toolkit (Bird 2006), and then tokenized, truecased, and normalised (in terms of punctuation) using Moses' scripts.

Participant profiles

Of the six participants (henceforth referred to as T1 to T6) in this experiment, four are male and two female. At the time of participation, three were aged 55-58, and three were aged 24-30.

[1] This back-translation used a purpose-built SMT system similar to the one previously described in this section, except that it was only trained using in-domain data and did not utilise continuous space *n*-grams.

When self-rating their proficiency of English using the CEFR[2] scale of language ability, three participants rated themselves at C2 (highly advanced), two at C1 (advanced), and one – the most experienced – at B2 (upper intermediate). One participant (T4) holds an undergraduate degree in English, and all others hold translation qualifications: three hold undergraduate translation and interpreting degrees, one with a Master's degree in Conference Interpreting; one participant holds a postgraduate qualification in literary translation (T3), and one holds a PhD (T6).

On average, participants have 13.2 years of professional translation experience (10.5 years in literary translation), although most have 3-10 years' experience and 1-4 years' literary translation experience. The two most experienced participants have 30 and 25 years' literary translation experience (T6 and T4), and have translated (circa) 100 and 17 novels, respectively. Other participants had translated one novel, 29 plays, and another literary translation (T2), two novels (T3), and one novel as part of a team (T5). The least-experienced participant (T1), who at the time of the experiment had translated a short film.

Only two participants had previous PE experience: one had post-edited for four years (T2) and one for one year (T3). T3 was the only respondent who agreed that they use and like MT. T5 would "only be willing to use it with very similar languages, with which it can offer good results". One participant uses MT "because I have to", and three agreed that "I don't use it and I'd prefer not to use it in the future". In their study in 2015, Moorkens and O'Brien found novice translators to have a more positive attitude to MT than experts. Similarly in this study, participants with 10 years' translation experience or less would mostly consider post-editing, regardless of age, whereas the two highly experienced translators do not use it and would prefer not to in the future.

Experimental Setup

Participants translated and post-edited the machine translated selections from the test data using the PET (Post-Editing Tool) interface (Aziz et al. 2012), an open source computer-assisted translation tool for editing at the segment level, built for research purposes. PET supports both translating from scratch and PE, and was used with its default settings. The six participants translated the first six sentences from the prologue of *Warbreaker* (two from scratch, two post-editing SMT, and two post-editing NMT) to familiarise themselves with the tool and workflow.

The source text used in the experiment is Chapter 1 of *Warbreaker*. It contains 330 sentences which were split into 33 translation tasks, each containing 10 sequential sentences. Segmentation was at the sentence level so as to record sentence- rather than task-level results of PE effort that could be normalised by word or character. The three experimental conditions were translation from scratch (HT), and PE the translation produced by systems SMT (shown in

[2] Common European Framework of Reference for Languages: Learning, Teaching, Assessment.

PET as MT1)[3] and NMT (MT2). Participants saw all conditions, but not all combinations, as they translated each job in one translation condition.

The task order (shown for the first seven of 33 tasks as an example in Table 1) was randomised, with the following three constraints:

1. For Task 1, translation condition was set to HT for translators T1 and T2, to MT1 for T3 and T4 and to MT2 for T5 and T6.
2. Two consecutive tasks by a translator cannot follow the same translation condition.
3. Each translator should complete an equal number of tasks under each translation condition (i.e. 11 HT tasks, 11 MT1 tasks, and 11 MT2 tasks).

	Translator					
Task ID	T1	T2	T3	T4	T5	T6
1 (first 10 sentences)	HT	HT	MT1	MT1	MT2	MT2
2 (sentences 11 to 20)	MT1	MT2	HT	HT	HT	HT
3	MT2	MT1	MT1	MT1	MT1	MT1
4	HT	HT	MT2	HT	MT2	MT2
5	MT1	MT2	MT1	MT1	MT1	MT1
6	MT2	MT1	HT	HT	MT2	HT
7	HT	HT	MT2	MT2	HT	MT1

Table 1. Task order for all participants for the first seven tasks

Prior to beginning the assigned tasks, participants were provided with comprehensive guidelines[4] in order to produce publishable professional quality translations, both for HT and for PE tasks. They were encouraged to retain the MT output where possible when translating, but could delete it and translate from scratch if they consider the quality of the MT output too low. PET aligns source and target segments on a one-to-one basis and displays them sequentially. However, participants could break that alignment, creating many-to-one (more than one source sentence translated as one target sentence) or one-to-many (one source sentence translated as more than one target sentence) translations where necessary.

Participants completed online pre- and post-task questionnaires[5] and recorded debriefing conversations with the authors at the end of the experiment wherein they further explained some questionnaire responses. These recordings were transcribed and coded, with findings

[3] Throughout the study, we referred to the two MT systems as MT1 and MT2 so that the translators could not know anything about the MT paradigm from which the translations emanated, in order to avoid any possible bias.

[4] Section 3 of https://github.com/antot/postediting_novel_frontiers/blob/master/pipenovel_translator_manual.pdf.

[5] Pre-task questionnaire is available at https://www.surveymonkey.com/r/TJCPWWM. Post-task questionnaire: https://www.surveymonkey.com/r/TFT56QW.

reported in the results section. Some participants also offered clarifications, explaining their post-task responses. The study was approved by the Research Ethics Committee of the Faculty of Arts at the University of Groningen. All participants received an explanation of the purpose of this research in plain language, gave their informed consent for their data (including timings, keystrokes, answers and recordings) to be used anonymously, and were paid an agreed per-word fee for their translation work and an extra flat-fee for compiling the pre- and post-questionnaires and having a debrief conversation.

Results

Measurements and Perceptions of Post-Editing Effort

Temporal effort

Figure 1 reports our measurement of temporal PE effort for each translator and translation condition in terms of the time required to translate the source text, measured in seconds, normalised by the number of characters in the segments of the source text that each translator translated under each translation condition.[6] In addition, the figure reports the perception of the translators, in terms of a ranked preference to the question “Rank the translation methods according to the translation speed”. Answers go from 1 (perceived the fastest) to 3 (perceived the slowest).

According to the measurements, all translators were faster post-editing SMT (condition MT1) than translating from scratch and they were all faster post-editing NMT (condition MT2) than SMT. The perceptions match the measurements in the case of NMT, as all the respondents perceived this condition to be the fastest (rank 1). The perceptions for SMT do not always match the measurements though; while all the translators were faster under this condition than translating from scratch (condition HT), two of them (T1 and T5) perceived themselves to be slower under this condition.

[6] Note (for this and other questions): T2 and T6 did not keep track of which system they were post-editing (MT1 or MT2), so we cannot analyse their perceptions for questions that ask for a comparison of both systems.

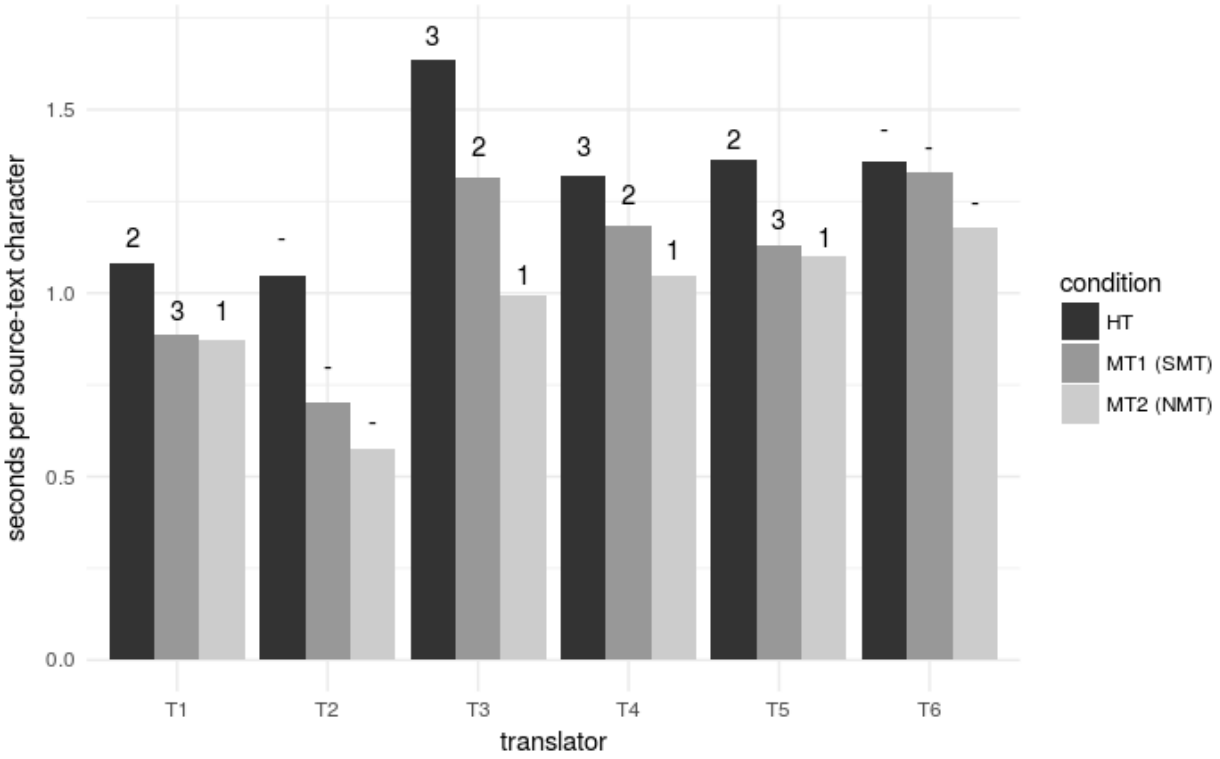

Figure 1. Measurements (displayed by bar height) and perceptions (displayed by rank) of temporal effort.

Technical effort

Figure 2 reports our measurement of PE technical effort for each translator and translation condition together with their perception, in terms of the number of keystrokes required to translate the source text, normalised by the number of characters in the source text. Their perception is again reported in terms of a ranked preference: from 1 (required the least effort) to 3 (required the most effort).

Comparing both MT systems, the perceptions match the measurements; all translators used fewer keystrokes when post-editing MT2 compared to post-editing MT1 and all of them reported post-editing MT2 requiring less effort than MT1.

When comparing HT to MT, however, the perceptions do not match the actual technical effort. T1 and T3 reported HT requiring less effort but they actually used more keystrokes for HT than for MT1 or MT2. Conversely, T4 and T5 reported that HT required the most effort, but under this condition they used fewer keystrokes than in condition MT1.

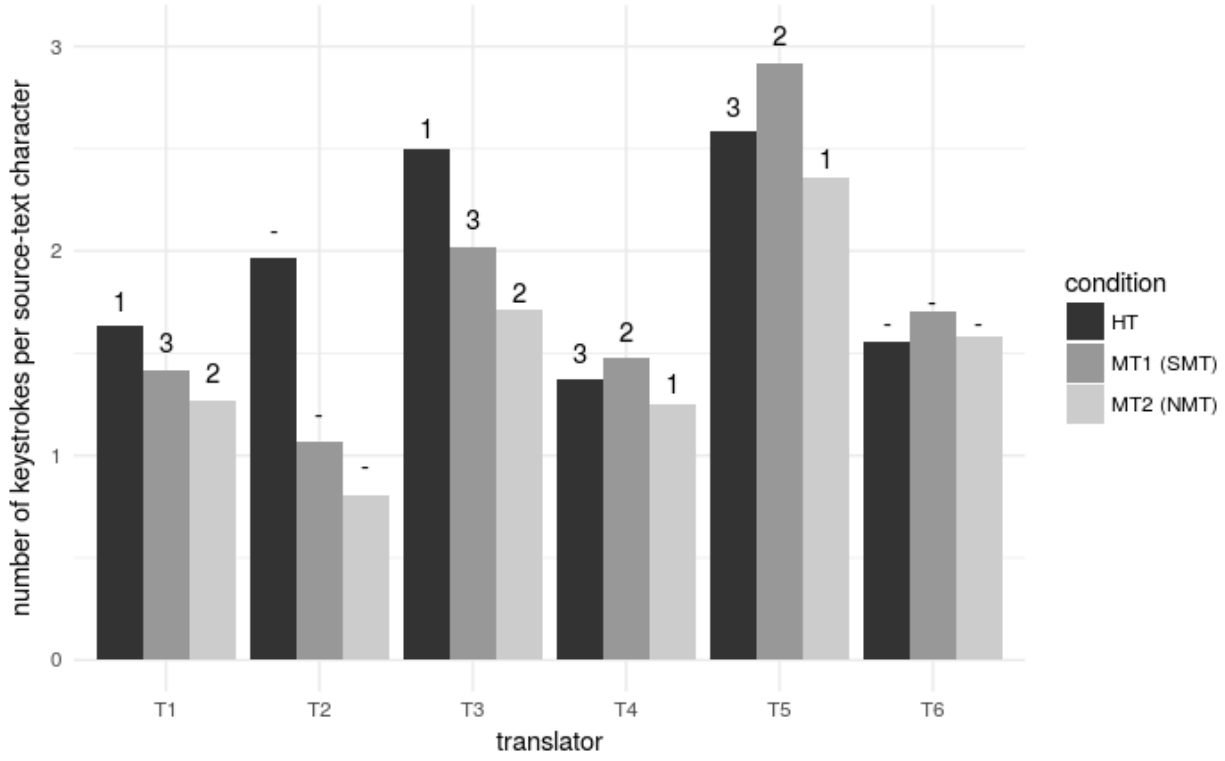

Figure 2. Measurements (displayed by bar height) and perceptions (displayed by rank) of technical effort.

Users' perceptions of the task

Creativity vs Productivity

Previous studies have shown that increases in temporal and technical effort are not always enough to convince translators to embrace PE, particularly when "they perceive the quality of the MT output to be inadequate for their purposes" (Cadwell, Castilho, O'Brien, and Mitchell 2016: 237). This finding is generally repeated here, as clarified by T5, who prefers NMT "if I'm thinking about productivity/revenue", but prefers HT "if my priority is to guarantee maximum quality". T4 echoes this view, explaining that they can produce "higher quality output (e.g. more creative solutions) [when] translating from scratch". T3 offers a cooking analogy, comparing pre-cooked (which would correspond to PE) and homemade food (which would correspond to translating from scratch): the former is more standard and "always tastes the same". This participant believes that, depending on the original author, the average bestseller reader would not notice the difference. Introducing the related idea of creative input, T3 suggests that "a reader that values creativity and style would notice."

Participants in a study carried out by Moorkens & O'Brien (2015: 79) complained of this "lack of creativity", and a "limited opportunity to create quality" when post-editing. In this study, T1 explained that, translating from scratch, "you structure yourself the way you want to translate", whereas when post-editing "you are limited: less creativity." As mentioned in the Introduction,

several participants discussed how they feel ‘primed’ by MT suggestions, whereby translators may be “inclined to respect it more than [they] should” (T6). “MT2 conditions you.” A similar sentiment was expressed by T5, who said that MT PE may restrict users to “a solution that is not the optimal.” T3 finds that “it makes you a bit lazy.” “You don’t feel like changing too many things”. This point has been made in the literature by Pym (2008: 8), who complained that translators using leveraging tools are “virtually obliged to accept the renditions that come from the client”, as well as Bowker (2007: 182), who noted that, when presented with a fully-formed target segment, it may be “difficult for the translator to think of a different way of expressing that notion”.

Some MT users feel positively about beginning a segment translation with an MT proposal. Cadwell et al (2016: 235) report from their focus groups with translators that “using MT gives them inspiration or ideas that they would not otherwise have and helps to ‘kick-start’ the translation process for them”. T2 in the current study, who has 4 years of PE experience, concurs, finding MT to help with some sentences. “It gives you a draft translation [that] you just need to fix and polish.” When asked which translation condition requires more effort, T2 chooses HT, “because with MT the first step (draft translation) is done. It also reduces effort as it gives you an interpretation of the source sentence. You avoid having to scan the English sentence to get its meaning, you can get it from the draft translation in your native language.” Some post-editors in the industry are encouraged to look at the target text first, a step that they may find counterintuitive, then to check the source text for adequacy, to avoid the content errors that have been found by Nitzke (2016) and others. It may be that T2, as the participant with the longest PE experience in the cohort, has been required to do this in the past.

Segmentation and Context

Segmentation at sentence level has become the norm for computer-aided translation (CAT), despite some criticism. Heyn (1998), for example, called it ‘peephole translation’ due to loss of context and internal coherence. Some translators try to mitigate this loss by segmenting at paragraph level, although this will mean loss of leverage. As CAT tools are not normally used for literary translation, participants find them restricting. Rather than attempting to create target segments with formal equivalence, translators of expressive texts (Reiss 1981) such as novels try to create what Nida (1964) called an ‘equivalent effect’. This may necessitate substitution with equivalent cultural references or idioms, and some redistribution of sections of the text, both of which are difficult or impossible within the restrictive environment of contemporary CAT tools.

T6, the most experienced participant, introduces this difficulty of translating without context, calling it “translation in the darkness”. “I feel a bit desperate about not having a bigger translation unit; at least a paragraph, or even better a whole page.” T6 explains this further as follows: “A translator has a global view of the text, MT has a fragmented view”. Similarly, “translating small units (sentence by sentence) means that I have a very fragmented vision of the overall translation”. For example, “the last translation unit in document 10 is ‘Stop that!’. Without [follow-on] context one cannot really know what it refers to”. T5 provides an example

scenario where context is important for an accurate translation of the word ‘child’ into Catalan; it can be translated as “fill” [son], “filla” [daughter], nen [child, masculine] or nena [child, feminine]. T2 has a similar criticism, explaining how “most of the time you still need to reformulate the whole target sentence; sometimes even the vocabulary choices are wrong because they are out of context, but most of the time you need to modify sentence structure or grammar almost from scratch.” For T2, a lack of context is one reason why they “only see the point of MT in very short, simple segments.” In this translation, T2 would have expected to have merged some sentences as longer segments are natural in Catalan, whereas MT outputs quite a literal translation of similar length to the source. For this reason, T2 says “HT is closer to the style of the target language, MT closer to the style of the source language.”

Source vs Target (adequacy vs fluency)

Despite the ‘attention’ paid to intrasentential context in NMT systems, participants still complained of overly literal MT outputs that focus on formal rather than dynamic equivalence (Nida 1964). Participants aim for literary translation that is “directed primarily toward equivalence of response rather than equivalence of form” (Nida 1964: 166). T6 explains that “my objective is to work in the fine details so the translation preserves the reading experience of the source.” Elsewhere, T6 explains that an empty text box provides “the freedom not to translate too close to the source, so one can find instead the equivalence to the intimate meaning of the source”

T2 points again to inherent grammatical differences between English and Catalan. “Catalan’s MT output followed usually the English structure, e.g. temporal information in the middle in English, but it should go usually in the beginning in Catalan. MT output would have it still in the middle, as in English.” This results in MT output that is a “calque of source language structures.” It is difficult to remove this from the final output, going back to the finding of being ‘primed by the MT output’. “With HT the reader reads your interpretation. With MT-assistance the machine interpretation is still there.”

SMT vs NMT

Participants’ reports of comparative MT quality tend to fit with what we have seen in previous studies as reported previously. T2, for example, finds sentence structure to be a weakness in SMT output, complaining of “calques from English structure”. For NMT, T3 cites “vocabulary” as a weakness, providing an example of where the word ‘guard’ was translated as ‘guardia’, whereas T3 considers that ‘centinela’ (meaning sentinel) would be more appropriate. T1 says that NMT is less literal and “better on complicated sentences, [with] better use of ‘pronoms febles’” (Catalan weak pronouns, similar to those in French).

T5 agrees that “NMT results in a much better translation [that] sounds quite natural”, although not always with the correct level of politeness (something that Sennrich, Haddow and Birch (2016c) have worked on, marking politeness as a side constraint). For SMT, T5 found “grammar and disambiguation problems.” “It got confused with nouns modifying other nouns and

translated them the [wrong] way around: the first would then be modified by the latter in the translated version." This results in some segments where the meaning was "completely twisted and changed." T4 found 'figurative language' and 'cultural items' predictably difficult for both systems.

Perception vs reality

Rating the MT systems from 1 to 5, where 1 was 'very unhelpful' and 5 was 'very helpful', participants gave the SMT output an average score of 3.75, with 4 for NMT. Participants also rated the outputs for fluency on a 1 to 4 scale, where 1 is 'incomprehensible' and 4 is 'flawless', giving SMT an average score of 2 and NMT 2.75. They rated adequacy by answering the question "How much of the meaning expressed in the source text is represented in the translation?" using a similar 1 to 4 scale, where 1 is 'none' and 4 is 'everything', giving SMT an average score of 2.5 and NMT a score of 3. These results are tabulated in Table 2.

MT Type	SMT	NMT
Helpful (1 to 5)	3.75	4
Fluent (1 to 4)	2	2.75
Accurate (1 to 4)	2.5	3

Table 2. Participant ratings of MT systems.

Despite translators rating post-editing NMT lower for temporal and technical effort than translating from scratch (cf. Figures 1 and 2, respectively), all participants preferred translating in the latter condition. T4 explained that "I feel I can reach higher quality output (e.g. more creative solutions) translating from scratch." When productivity results were given to participants, highlighting the percentage of time saved, they still preferred HT. T3 gave a nuanced response when informed that "post-editing MT2, compared to from scratch, saved you 32% in term of time and 28% in terms of keystrokes." We asked whether, knowing this, "would it change your preference of translation method?" T3 responded: "If you're in a hurry, or if the text is repetitive, then NMT [will be useful]. If you want to do a more creative job and you've time, I keep my preference for HT."

Alternative Methods of Machine Assistance

Translators were asked to express their opinion on two alternative ways in which they could receive assistance from the MT system. Specifically, we asked them whether these alternatives could be appropriate and/or useful for translating literary text.

The first alternative regards quality estimation (Specia and Shah 2018), a method that, given a source segment and an MT system, predicts the quality of the MT output, for example in terms of its usefulness to be post-edited. The question assumes the existence of an accurate quality estimation system and was phrased as follows:

> If there was a mixed translation method in which you would post-edit MT only when the MT is good enough and translate from scratch complex sentences for which the MT output is bad, would you prefer that over translating from scratch?

T2, T3 and T5 thought that it would be helpful. T4 thought that "it would be fantastic in terms of optimising time, but in terms of quality still felt that translating from scratch is better". T6 did not find this alternative helpful as it would disturb the translation rhythm: "I like to translate the easy parts from scratch too, it is useful as a time in which I can rest after more difficult parts".

The second alternative concerned interactive MT, an alternative to post-editing in which translators work as they would when translating from scratch (i.e. without MT assistance), with the difference that, as they progress through a text, they receive autocomplete suggestions from an MT engine (Green, Heer, and Manning 2013). Thus, while PE is sequential and machine-guided, interactive MT is human-guided. PE, as noted previously, has been shown to prime the translator and limit creativity. Conversely, it is hypothesised that interactive MT does not limit the translator, as she is the one guiding the translation. Translators were asked this question:

> If there was a translation method in which you would translate from scratch but as you do so you would receive auto-complete suggestions (e.g. the next word or even the next 5 to 10 words), do you think that that could be a better method than translating from scratch?

T1 did not think this is a good idea: "If a suggestion is good but not exactly the same as the idea in your head that leads to high cognitive effort". T2 expressed a preference for PE: "PE is better because it does the first step for you [providing you with a draft translation to work upon]. Interactive MT would still require the translator to do this first step". T3 and T6 were skeptical but still open to it, e.g. T3 answered "I don't like it in principle, but I could get used to it". T4 and T5 were enthusiastic about this type of machine assistance.

Conclusion

This experiment set two MT systems the challenging task of translating English-to-Catalan literary texts, known to be difficult for automatic translation. Statistical and neural MT systems were built, trained on the same in-domain literary data, and six participants with experience of literary translation tried translating sections of the text from scratch, post-editing SMT, and post-editing NMT (in a random order), to investigate the effort required and their perceptions of the task.

All participants were faster when post-editing NMT, but they all still stated a preference for translation from scratch, as they felt less constrained and could be more creative. They complained that the MT systems 'conditioned' them to produce a literal translation, and found

the limitation of sentential segmentation awkward. When comparing MT systems, participants found NMT output to be more fluent and adequate. Both systems had trouble with ambiguity and mistranslation, and SMT output had further structural problems, often producing ‘calques’ of English. Participants considered that NMT could be useful if they need to produce a translation in a hurry, but they would not post-edit by choice.

Our findings show that, for this study, text type, and language pair, the move to NMT demonstrates an improvement in terms of productivity and number of edits required. Bar-Hillel (1960: 136) wrote that, without “extra-linguistic knowledge”, a translation system would be “in no position to resolve semantic ambiguities”. The ability to make calculations based on intrasentential context means that NMT systems can sometimes translate ambiguous words and phrases correctly where prior systems would have failed. Intersentential context may be the next step for NMT. However, participants in this study identified broader knowledge required to create an equivalent reading experience.

Comparing human translation ability to a non-conscious algorithm is problematic, as contemporary MT systems merely attempt to replicate the patterns found in their training data. The term Machine *Translation* may be considered something of a misnomer when the limit remains what Catford (1965) called ‘transference’, rather than translation, with the caveat that attributing even this ability to a non-conscious algorithm is problematic, as the system merely attempts to replicate the patterns found in its training data. Despite claims of NMT systems achieving human parity in certain language pairs and domains (Hassan et al. 2018), when compared with the competences expected of an expert human translator, replicating strategic competence and extra-linguistic competence does not appear close (PACTE 2005). The recent improvements in MT quality, along with the hype generated by reports of NMT quality (Castilho et al. 2017b), have been unnerving for some translators. However, experienced translators who become familiar with NMT output are more likely to identify its limitations. Participant T5 in this study says that it is “scary and frightening to see that the machines are getting better”, but nonetheless concludes that MT systems are “still far [from being a threat]”.

References

Aziz, Wilker, Sheila Castilho M. de Sousa, and Lucia Specia. 2012. “PET: a tool for post-editing and assessing machine translation.” In The Eighth International Conference on Language Resources and Evaluation, LREC′12, Istanbul, Turkey. May 2012.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. “Neural Machine Translation by Jointly Learning to Align and Translate”. arXiv preprint, arXiv:1409.0473.

Bar-Hillel, Yehoshua. 1960. “The Present Status of Automatic Translation of Languages.” Advances in Computers 1: 91-163.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, Marcello Federico. 2017. Neural versus Phrase-Based MT Quality: an In-Depth Analysis on English-German and English-French. *Computer Speech & Language* 49: 52-70.

Besacier, Laurent. 2014. “Traduction automatisée d’une oeuvre littéraire: une étude pilote.” In: Traitement Automatique du Langage Naturel (TALN), Marseille, France.

Bowker, Lynne 2007 “Translation Memory and ‘‘Text’’. In *Lexicography, terminology, and translation. Text- based studies in honour of Ingrid Meyer,* edited by Lynne Bowker, 175- 187, Ottawa: University of Ottawa Press.

Cadwell, Patrick, Sheila Castilho, Sharon O’Brien, and Linda Mitchell 2016. “Human factors in machine translation and post-editing among institutional translators.” *Translation Spaces* 5(2): 222-243.

Cadwell, Patrick, Sharon O’Brien, and Carlos S. C. Teixeira. 2017. “Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators.” Perspectives (online first).

Carl, Michael, Silke Gutermuth and Silvia Hansen-Schirra. 2015. “Post-editing machine translation: Efficiency, strategies, and revision processes in professional translation settings.” In *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting*, edited by Aline Ferreira and John W. Schwieter, 145–174, Amsterdam: John Benjamins.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli Barone, Maria Gialama. 2017a. “A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators.” MT Summit 2017, Nagoya, Japan.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, Andy Way. 2017b. “Is Neural Machine Translation the New State of the Art?” *The Prague Bulletin of Mathematical Linguistics* 108: 109-120.

Castilho, Sheila, Sharon O'Brien. 2016. “Content profiling and translation scenarios.” *The Journal of Internationalization and Localization* 3(1): 18-37.

Catford, John C. 1965. *A Linguistic Theory of Translation: An Essay in Applied Linguistics*. London: Oxford University Press.

Church, Kenneth W., Eduard H. Hovy. 1993. “Good applications for crummy machine translation.” Machine Translation 8(4): 239–258.

Daems, Joke, Orphée De Clercq, Lieve Macken. 2017. “Translationese and post-editese: How comparable is comparable quality?” Linguistica Antverpiensia, New Series: Themes in Translation Studies 16: 89–103.

De Almeida, Giselle, Sharon O'Brien. 2010. "Analysing post-editing performance: Correlations with years of translation experience." In Proceedings of the 14th annual conference of the European Association for Machine Translation, St. Raphaël, France.

Durrani, Nadir, Helmut Schmid, Alexander Fraser. 2011. "A joint sequence translation model with integrated reordering," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies- Vol. 1, Portland, OR, 1045–1054.

Forcada, Mikel. 2017. "Making sense of neural machine translation." Translation Spaces 6(2): 291-309.

Gaspari, Federico, Antonio Toral, Sudip Kumar Naskar, Declan Groves, Andy Way. 2014. "Perception vs Reality: Measuring Machine Translation Post-Editing Productivity." In Proceedings of AMTA 2014 Workshop on Post-editing Technology and Practice, Vancouver, 60-72.

Genzel, Dmitriy, Jakob Uszkoreit, Franz Och. 2010. "'Poetic' Statistical Machine Translation: Rhyme and Meter." In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, 158–166.

Green, Spence, Jeffrey Heer, Christopher D. Manning. 2013. "The Efficacy of Human Post-Editing for Language Translation." ACM Human Factors in Computing Systems (CHI).

Greene, Erica, Tugba Bodrumlu, Kevin Knight. 2010. "Automatic analysis of rhythmic poetry with applications to generation and translation." In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, 524–533

Guerberof, Ana. 2012. "Productivity and quality in the post-editing of outputs from translation memories and machine translation." PhD Dissertation. Universitat Rovira i Virgili.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. "Achieving Human Parity on Automatic Chinese to English News Translation." Redmond: Microsoft AI & Research.

Heyn Matthias 1998. "Translation memories: Insights and prospects." In *Unity in diversity? Current trends in translation studies*, edited by Lynne Bowker, Michael Cronin, Dorothy Kenny, and Jennifer Pearson, 123-36. Manchester: St. Jerome.

Jones, Ruth, Ann Irvine. 2013. "The (un)faithful machine translator." In Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Sofia, Bulgaria, 96–101.

Klubička, Filip, Antonio Toral, Víctor M. Sánchez-Cartagena. 2017. “Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation.” *The Prague Bulletin of Mathematical Linguistics* 108: 121-132.

Koehn P, Knowles R (2017) Six Challenges for Neural Machine Translation. In: Proceedings of the First Workshop on Neural Machine Translation, Vancouver, BC, Canada, pp 28–39

Koponen, Maarit. 2016. Is post-editing worth the effort? A survey of research into post-editing and effort. *Journal of Specialised Translation* 25: 131-148.

Krings, Hans P. 2001. Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes. Ohio: Kent State University Press.

Lacruz, Isabel and Gregory M. Shreve. 2014. “Pauses and Cognitive Effort in Post-Editing.” In Post-editing of Machine Translation: Processes and Applications edited by Sharon O’Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia, 287-314, Newcastle-Upon-Tyne: Cambridge Scholars.

LeBlanc, Matthieu (2013). “Translators on translation memory (TM). Results of an ethnographic study in three translation services and agencies.” *The International Journal for Translation and Interpreting Research* 5(2): 1-13.

Lommel, Arle, and Donald A. DePalma. 2016. “Europe’s Leading Role in Machine Translation: How Europe Is Driving the Shift to MT.” Boston: Common Sense Advisory.

Martín, Juan Alberto Alonso, and Anna Civil Serra. 2014. “Integration of a Machine Translation System into the Editorial Process Flow of a Daily Newspaper.” *Procesamiento del Lenguaje Natural, Revista* 53: 193-196.

Moorkens, Joss, Sharon O’Brien. 2015. “Post-Editing Evaluations: Trade-offs between Novice and Professional Participants.” In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT 2015),* edited by İlknur Durgar El- Kahlout, Mehmed Özkan, Felipe Sánchez- Martínez, Gema Ramírez- Sánchez, Fred Hollowood, and Andy Way, 75-81.

Moorkens, J., Sharon O’Brien, Igor A. L. Silva, Norma Fonseca, Fabio Alves. 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, Joss, Sharon O’Brien. 2017. “Assessing User Interface Needs of Post-Editors of Machine Translation.” In Human Issues in Translation Technology: The IATIS Yearbook, edited by Dorothy Kenny, 109-130, Oxford, UK: Routledge.

Moorkens, Joss. 2018. “Eye-Tracking as a Measure of Cognitive Effort for Post-Editing of Machine Translation.” In Eye Tracking and Multidisciplinary Studies on Translation, edited by Callum Walker and Federico Federici, 55-69, Amsterdam: John Benjamins.

Nida, Eugene. 1964. *Towards a Science of Translating*. Leiden: Brill.

Nitzke, Jean. 2016. “Monolingual post-editing: An exploratory study on research behaviour and target text quality.” In Eye-tracking and Applied Linguistics, edited by Silvia Hansen-Schirra & Sambor Grucza, 83–109. Berlin: Language Science Press.

PACTE Group. 2005. “Investigating Translation Competence: Conceptual and Methodological Issues, *Meta* 50(2): 609-619.

Plitt, Mirko, François Masselot. 2010. “A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context.” *The Prague Bulletin of Mathematical Linguistics* 93: 7–16.

Pym, Anthony. 2008. “Professional corpora: teaching strategies for work with online documentation, translation memories and content management.” *Chinese Translator’s Journal*, 29(2): 41- 45.

Reiss, Katharina. 1981. “Type, Kind and Individuality of Text: Decision Making in Translation.” *Poetics Today* 2(4): 121-131.

Sennrich, Rico, Barry Haddow, Alexandra Birch. 2016a. “Improving Neural Machine Translation Models with Monolingual Data.” In proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 86–96, Berlin, Germany, August 7-12, 2016.

Sennrich, Rico, Barry Haddow, Alexandra Birch. 2016b. “Neural Machine Translation of Rare Words with Subword Units.” In proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1715-1725, Berlin, Germany, August 7-12, 2016.

Sennrich, Rico, Barry Haddow, Alexandra Birch. 2016c. “Controlling Politeness in Neural Machine Translation via Side Constraints.” In proceedings of NAACL-HLT 2016, 35–40.

Sennrich, Rico, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, Maria Nadejde. 2017 Nematus: a Toolkit for Neural Machine Translation. In Proceedings of the Software Demonstrations from the 15th Conference of the European Chapter of the Association for Computational Linguistics, 65-68.

Somers, Harold. 2001. *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamins.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. “A Study of Translation Edit Rate with Targeted Human Annotation.” Proceedings of Association for Machine Translation in the Americas.

Specia, Lucia. 2011. “Exploiting Objective Annotations for Measuring Translation Post-editing Effort.” In proceedings of the 15th Conference of the European Association for Machine Translation, 73–80, Leuven, Belgium.

Specia, Lucia, and Kashif Shah. 2018 “Machine Translation Quality Estimation: Applications and Future Perspectives.” In *Translation Quality Assessment: From Principles to Practice,* edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, 201-236, Heidelberg: Springer.

Teixeira, Carlos S. C. 2014. “Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories.” In Proceedings of the Third Workshop on Post-Editing Technology and Practice (WPTP-3), edited by Sharon O’Brien, Michel Simard, and Lucia Specia, 45–59.

Thouin, Benoît. 1982. “The Meteo system.” In Practical Experience of Machine Translation: Proceedings of Translating and the Computer 1981, edited by Veronica Lawson, 39-44, Amsterdam: North-Holland.

Toral, Antonio, Andy Way. 2015. Translating Literary Text between Related Languages using SMT. In proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature, pages 123–132, Denver, Colorado.

Toral. Antonio, Victor M. Sánchez-Cartagena. 2017. “A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions.” In Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017. Valencia, Spain.

Toral, Antonio, Andy Way. 2018 “What level of quality can Neural Machine Translation attain on literary text?” In *Translation Quality Assessment: From Principles to Practice,* edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, 263-287, Heidelberg: Springer.

Toral, Antonio, Martijn Wieling, and Andy Way. 2018. “Post-editing Effort of a Novel with Statistical and Neural Machine Translation”. *Frontiers in Digital Humanities* 5:9. doi: 10.3389/fdigh.2018.00009

Vasconcellos, Muriel. 1985. “Machine aids to translation: a holistic scenario for maximizing the technology.” In *Overcoming language barriers: the human/machine relationship,* edited by Humphrey Tonkin and Karen Johnson-Weiner, 27-34. Report of the Fourth Annual Conference of the Center for Research and Documentation on World Problems, New York, December 13-14.

Vaswani, Ashish, Yinggong Zhao, Victoria Fossum and David Chiang. (2013). “Decoding with large-scale neural language models improves translation”, in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, 1387–1392.

Viera, Lucas Nunes. 2014. “Indices of cognitive effort in machine translation post-editing.” *Machine Translation* 28(3-4):187-216.

Way, Andy. 2013. “Traditional and emerging use-cases for machine translation.” In Proceedings of Translating and the Computer 35, London, UK.

Way, Andy. 2018a. “Quality expectations of machine translation.” In *Translation Quality Assessment: From Principles to Practice,* edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, 159-178, Heidelberg: Springer.

Way, Andy. 2018b. Machine Translation: Where we are at today. In *The Bloomsbury Companion to Language Industry Studies*, edited by Erik Angelone, Gary Massey, and Maureen Ehrensberger-Dow, London: Bloomsbury.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017 “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. arXiv preprint 1609.08144 (https://arxiv.org/abs/1609.08144).