Human evaluation of three machine translation systems: from quality to attitudes by professional translators

Anna Fernández-Torné
Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental
Universitat Autònoma de Barcelona
anna.fernandez.torne@uab.cat

Anna Matamala
Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental
Universitat Autònoma de Barcelona
anna.matamala@uab.cat

Abstract

This article aims to compare three machine translation systems with a focus on human evaluation. The systems under analysis are a domain-adapted statistical machine translation system, a domain-adapted neural machine translation system and a generic machine translation system. The comparison is carried out on translation from Spanish into German with industrial documentation of machine tool components and processes. The focus is on the human evaluation of the machine translation output, specifically on: fluency, adequacy and ranking at the segment level; fluency, adequacy, need for post-editing, ease of post-editing, and mental effort required in post-editing at the document level; productivity (post-editing speed and post-editing effort) and attitudes. Emphasis is placed on human factors in the evaluation process.

Keywords: machine translation, quality evaluation, human evaluation, automatic metrics, post-editing effort

Resumen

En este artículo se comparan tres sistemas de traducción automática poniendo especial atención en la evaluación humana. Los sistemas analizados son un sistema estadístico de traducción automática con adaptación al dominio, un sistema neuronal de traducción automática con adaptación al dominio y un sistema de traducción automática genérico. La comparación se lleva a cabo en una traducción del español al alemán de documentación industrial de componentes y procesos de máquina herramienta. El estudio se centra en la evaluación humana de la traducción automática, en concreto en los siguientes aspectos: fluidez, adecuación y ranquin a nivel de segmento; fluidez, adecuación, necesidad de posedición, facilidad de posedición

y esfuerzo mental requerido en la posedición a nivel de documento; productividad (velocidad de posedición y esfuerzo de posedición) y actitudes. Se hace énfasis en los factores humanos del proceso de evaluación.

Palabras clave: traducción automática, evaluación de la calidad, evaluación humana, métricas automáticas, esfuerzo de posedición

1. Introduction

Machine translation research has seen two interesting developments in recent years: firstly, the rise of neural machine translation (Cho et al., 2014; Castilho et al., 2017) and, secondly, the willingness to go beyond automated metrics such as BLEU (Papineni et al., 2002) or TER (Snover et al., 2006) and seek feedback from human participants (Callison-Burch et al., 2012). Our research combines the two approaches and aims to compare three different machine translation systems with a focus on human evaluation in a domain-specific scenario where resources are scarce. Making proper domain adaptation is not only a goal but also a challenge which has been researched extensively within statistical machine translation (SMT) (Foster & Kuhn, 2007, Axelrod et al., 2011; Bisazza et al., 2011; Gascó et al., 2012; Sennrich, 2012; Eetemadi et al., 2015), but to a lesser extent in neural machine translation (NMT) (Luong & Manning, 2015; Freitag & Al-Onaizan, 2016; Crego et al., 2016). Large generic machine translation systems are freely available online and are used even in cases where domain-adapted systems may be more suitable, but evaluations comparing large generic MT systems and domain-adapted systems, with an emphasis on human evaluation, are missing. The systems under analysis in our research are a domain-adapted statistical machine translation (SMT) system, a domain-adapted neural machine translation (NMT) system and a generic machine translation (GT) system. They have been compared in three domains and language pairs: reports and press releases from non-profit international organisations (from English into Spanish) (INTORG), industrial documentation of machine tool components and processes (MTOOL) (from Spanish into German), and the installation and maintenance documentation for elevators (ELEV) (from Spanish into French). Automated metrics have been computed and the global results have already been presented (Etchegoyhen et al., 2018), but the aim of this article is to focus on the results of just one language pair and domain, namely MTOOL (from Spanish into German), in order to provide a more thorough discussion, with a focus on human factors that were not previously discussed.

Section 2 describes the corpora and models used. Section 3 describes methodological aspects, i.e. the measures used in the human evaluation, the tool

selected to perform the experiment, the participants' selection procedure and profile, and the development of the test. Section 4 discusses the results of the MTOOL evaluation. The article concludes with some thoughts on future research avenues concerning human factors in machine translation research.

2. Corpora and MT models in MTOOL

This research was developed as part of the AdapTA project, in which data were obtained from project partners. For the MTOOL (from Spanish into German) corpus, the training data provided by the partner specialised in the domain were particularly scarce. Thus, only 25,256 parallel segments were gathered, 1,984 segments were used as development sets, and three sets of 50 sentence pairs were selected for testing purposes. The selection was not random but took into account three factors to guarantee that sentences were representative and could be used in tasks replicating a real professional scenario: the presence of specific domain vocabulary, the average sentence length (as in any technical field, sentences are mostly short in this domain too), and the presence of coherent segments at the contextual level. The content was highly specialised and dealt with industrial documentation of machine tool components and processes. This scenario replicates a typical situation for which there is a high demand from a professional point of view and limited training resources. To complement the scarce data provided in the form of translation memories for MTOOL, out-of-domain data were compiled, with a total of 1,784,385 additional segments obtained from freely available corpora (see Etchegoyhen et al., 2018 for further technical details). The main function of these generic datasets was to serve as a basis for the NMT models.

As explained by Etchegoyhen et al. (2018), SMT systems were phrase-based models built with Moses (Koehn et al. 2007), with phrases of maximum length 5 and n-gram language models of order 5 built with KenLM (Heafield, 2011). For NMT, the attention-based encoder-decoder approach (Bahdanau et al., 2015) was followed, using the OpenNMT toolkit (Klein et al., 2017). The translations from the online generic system were obtained from Google Translate in June 2017 in which, to the best of our knowledge, translations from Spanish into German were produced using their phrase-based SMT engine. Domain adaptation in MTOOL was carried out using the system that performed best during tests: for SMT, the phrases from the entire generic dataset were combined through fill-up and, for NMT, it was carried out through fine-tuning (Luong & Manning, 2015), by training the generic networks on the in-domain data.

3. Methodological aspects

A field quasi-experiment, i.e., one "taking place in real life [...] in which the criterion of randomization (of participants in a sample, for instance) cannot be met" (Van Peer, Hakemulder and Zyngier, 2012: 90) was planned. Thus, a more realistic but less controlled environment was favoured, prioritising its ecological validity since "a laboratory often fails to replicate the everyday conditions under which cultural phenomena occur" (ibid: 89). The aim was to gather both qualitative and quantitative data. The experiment was approved by UAB's Ethical Committee on Animal and Human Research (CEEAH) and, following the committee's advice, the experiment included one part in which the participants were paid their requested fees as professional translators to carry out a series of tasks and another part (replying to questionnaires) which was voluntary.

3.1. Selecting the measures for human evaluation

The human evaluation took into account three factors: quality at the segment and document levels, productivity, and translators' attitudes. At the segment level three indicators were gathered to assess quality: fluency, adequacy and ranking. Fluency is understood to be the extent to which a translated segment flows naturally in the target language without grammar and spelling mistakes and is considered to be genuine language by native speakers (Koehn and Monz, 2006). It was measured on a 1 to 4 scale, with 1 indicating that the text was incomprehensible and 4 indicating that the text was flawless, following TAUS guidelines. Adequacy, measured on another 4-point scale, with 1 indicating none of the meaning is represented in the translation and 4 indicating everything is represented in the translation, is considered to be the amount of information from the original segment that is present in the translated segment (Koponen, 2010). As regards ranking, it consists of placing in order different translated versions from the same original segment from best (1) to worst quality (3).

At the document level, the subjective perception of participants regarding five quality aspects was also gathered, namely fluency, adequacy, need for post-editing, ease of post-editing, and mental effort involved in the post-editing. Those five aspects were rated on a 10-point scale and were presented to the participants as follows, with not specific definition added:

- How fluent the raw machine translated text was, with 1 indicating that the text was not fluent and 10 indicating that it was very fluent.

- How much of the information in the source text was present in the raw machine translated text, with 1 indicating none of the information in

the source text was represented in the translation and 10 indicating all information in the source text was represented in the translation.

- How much post-editing the text required, with 1 indicating the translation required very few editing and 10 indicating the translation required a lot of editing.

- How easy the post-editing was, with 1 indicating the post-editing was found very difficult and 10 indicating the post-editing was found very easy.

- How much mental effort the post-editing required, with 1 indicating the post-editing required very low mental effort and 10 indicating the post-editing required a lot of mental effort.

They were also given the opportunity to add comments after each statement. Concerning productivity, the objective measures chosen were post-editing speed and post-editing effort. Post-editing speed refers to the "average number of words processed by the post-editor in a given timespan" (TAUS, n.d.), measured in words per hour. Post-editing effort is defined as "the average percentage of word changes applied by the post-editor on the MT output provided" (TAUS, n.d.). The effort is measured on a 0 to 10 scale in which 0 means that no changes needed to be made on the MT output and 10 implies that all the text or most of it was changed. It is based on the edit distance (Levenshtein's algorithm) normalised by segment length (i.e., divided by the number of characters of whatever segment is longer: either the automatically translated one or the post-edited one).

As regards attitudes, our aim was to assess the attitude of the participants prior to the test and its evolution during the experiment through a questionnaire administered before and after the three tasks. The questionnaire included a 1-to-10 scale in which participants gave their opinion on the following aspects:

- general machine translation quality (without post-editing), with 1 indicating the raw MT is of very poor quality and 10 indicating it is a very good raw MT.

- usefulness of machine translation for translators, with 1 indicating that MT is useless and 10 indicating it is very useful.

- inclination to use machine translation as a starting point, with 1 indicating a very low inclination and 10 indicating a very high inclination to use MT.

- interest in post-editing, with 1 indicating a very low interest and 10 indicating a very high interest in the use of MT.

- boredom of post-editing tasks, with 1 indicating that post-editing tasks are not boring at all and 10 indicating they are very boring.

- cognitive effort involved in post-editing tasks, with 1 indicating a very low cognitive effort and 10 indicating a very high cognitive effort involved in post-editing tasks.

- and quality of post-edited machine translated texts, with 1 indicating the post-edited MT texts are of very poor quality and 10 indicating they are very good post-edited machine translated texts.

3.2. Selecting the tool

After analysing various tools such as CASMACAT, MATECAT, PET, Translog II, Appraise, Costa MT, MT-Equal, TransCenter and CATaLog Online, TAUS DQF was chosen due to its extensive use both in academia and in the industry (Görög, 2014; Valli, 2015). It has a user-friendly interface that makes the process easier both for the researcher and the translator. TAUS DQF also generates reports of the results automatically and randomises the presentation of segments in the ranking task. However, the current version of the tool has two drawbacks, i.e. it does not allow the post-editor to have a global view of the text and it does not allow the post-editor to go back to previous segments (Moran, Saam @ Lewis, 2014).

3.3. Selecting the participants

A priori non-probabilistic purposive sampling and snowball sampling techniques were used for the recruiting of respondents (Bryman, 2012) according to the following criteria: they should be professional translators working in the language pair being researched and native speakers of the target language. They were identified through distribution lists and email contacts. In MTOOL (Spanish into German), 22 professionals participated in the experiment, but due to technical issues only socio-demographic data from 21 were recorded. Seventeen were women (81%) and the age range was between 32 and 67. Most participants (95%) had university education and they all had at least 2 years' experience in translation. 19 translators (91%) had experience in the revision of third-party texts and 11 (52%) had also worked as professional post-editors.

3.4. Test development

The test lasted 4 hours approximately and, to avoid participants' fatigue, it was divided into two sessions that participants undertook at their own convenience in

a 3-week period. First of all, participants were informed via email of the tasks to be carried out in the first session. After signing the informed consent sheet, they were asked to participate in a voluntary section in which a general questionnaire collected socio-demographic data, followed by a post-editing pre-questionnaire (see Section 3.1).

Next, they were asked to post-edit three texts (containing 50 segments each with a total of 843, 985 and 953 words) for which they were paid a fee. The order of the presentation of texts and the MT system used to translate such texts was randomised and each individual participant given instructions in a specific order. After post-editing each text, participants were requested to assess fluency, adequacy, need for post-editing, ease of post-editing, and mental effort involved in the post-editing on a 1 to 10 scale (see Section 3.1).They were also given the opportunity to add comments after each statement.Finally, they were requested to reply again to the same questions as in the pre-questionnaire on post-editing, to see how much their attitude had changed after the post-editing task.

Once they finished the first part, participants received a second e-mail with the instructions for the second session, in which they had to carry out different tasks on the same 150 segments: a fluency evaluation task, an adequacy assessment task, and a ranking task.

3.5. Automated metrics

Automated metrics were computed for the three systems. Table 1 summarises the values and shows the positive impact of domain adaptation both in SMT and NMT in contrast with a generic system, which is especially relevant taking into account the limited amount of training data when compared to generic systems. It also shows how fine-tuned NMT systems seem to perform better than the other ones. For the BLEU metric statistical significance at $p < 0.05$ was found between all pairs of systems, i.e., between NMT and GT, between NMT and SMT, and between GT and SMT.

Table 1: Objective automated metrics

	SMT	NMT	GT
BLEU	19.830	27.715	12.265
METEOR	35.260	41.471	25.668
TER	69.378	62.203	85.055

4. Discussion of results

Results from the human evaluation will be presented separately for each element assessed. A statistical analysis was carried out using IBM SPSS v.20, with a significance level of 0.05.

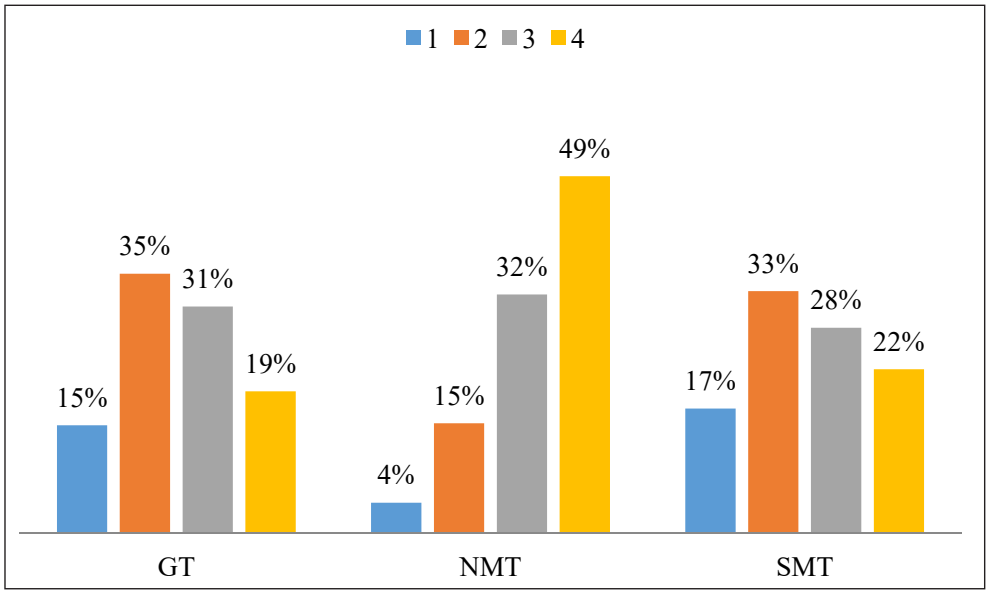
Chi-square tests (Saldanha & O'Brien, 2013) were employed to compare the distribution of qualitative data, i.e., the adequacy, fluency and ranking assessments at the segment level. Discrete quantitative data, such as quality assessments at the textual level, were analysed using a Mann-Whitney U test (Mann and Whitney, 1947) to compare groups, while for continuous quantitative data, such as PE speed and effort, the Bonferroni-corrected Mann-Whitney U test was used for multiple comparisons (Dunn, 1964). Attitude assessments were considered as paired discrete numerical variables and a Wilcoxon signed-rank test (Wilcoxon, 1945) was then applied to determine whether there was a statistically significant change in their opinions. Moreover, Spearman's correlation (Schober, Boer & Schwarte, 2018) was used to see if socio-demographic data and participants' professional experience had any influence on the different assessments.

An inter-rater reliability analysis was also performed using the quality assessment variables at the segment level through the intra-class correlation coefficient (ICC) estimates. Thus, the ranking obtained an ICC of 0.924, reaching 0.956 in the case of the adequacy and 0.979 in the case of the fluency, which are excellent levels of reliability.

4.1. Quality at the segment level

In terms of adequacy, assessed on a 1 to 4 scale, the GT and SMT systems show a similar distribution, with most segments being assessed as a 2 and a 3 (35% and 31% in GT; 33% and 28% in SMT, respectively), whilst in the NMT system participants rate most segments with higher marks, 3 (32%) and 4 (49%), as shown in Graph 1.

Graph 1: Adequacy metrics



The median for all three systems is 3, but the mode for the NMT system is 4, much higher than for the GT and SMT systems (mode = 2). The same differences are found when mean rates are compared, as the NMT system shows a mean of 3.25 (*SD* = 0.86) and the GT and SMT systems show similar lower mean values (M = 2.55, *SD* = 0.96 and M = 2.56, *SD* = 1.02 respectively), as Table 2 shows:

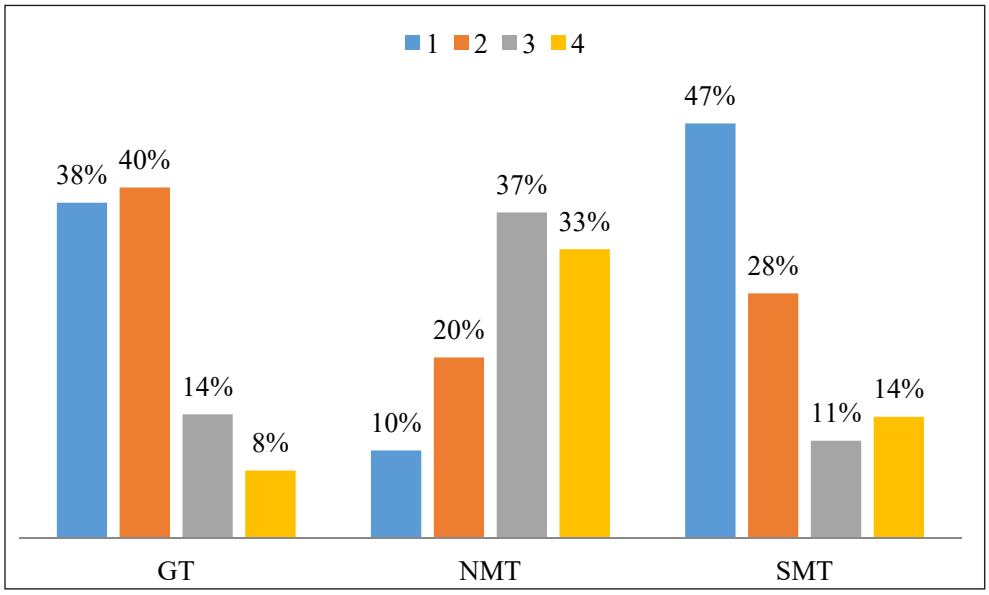
Table 2: Descriptive statistics for adequacy

	NMT	GT	SMT
Mean	3.25	2.55	2.56
Median	3	3	3
Mode	4	2	2

Differences between GT and SMT are not statistically significant, but they are between GT and NMT ($X^2(3) = 296.28$, $p < 0.001$; $Z = 510.5$, $p < 0.05$) and between NMT and SMT ($X^2(3) = 271.48$, $p < 0.001$; $Z = 1,896$, $p < 0.05$), hence proving that the NMT system is the best rated in terms of adequacy.

In terms of fluency, in the GT and SMT systems more than 75% of the segments are rated with low values, as shown in Graph 2. On the contrary, the NMT system only has 30% of segments in this low range, showing an entirely different pattern.

Graph 2: Fluency metrics



All descriptive values, included in Table 3, show better values for the NMT system, followed by the GT and SMT systems. The mean value for the NMT system (M = 2.92, SD = 0.96) is higher than that of the GT and SMT systems, although the difference between the GT system (M = 1.906, SD = 0.91) and the SMT system (M = 1.909, SD = 1.06) is almost non-existent.

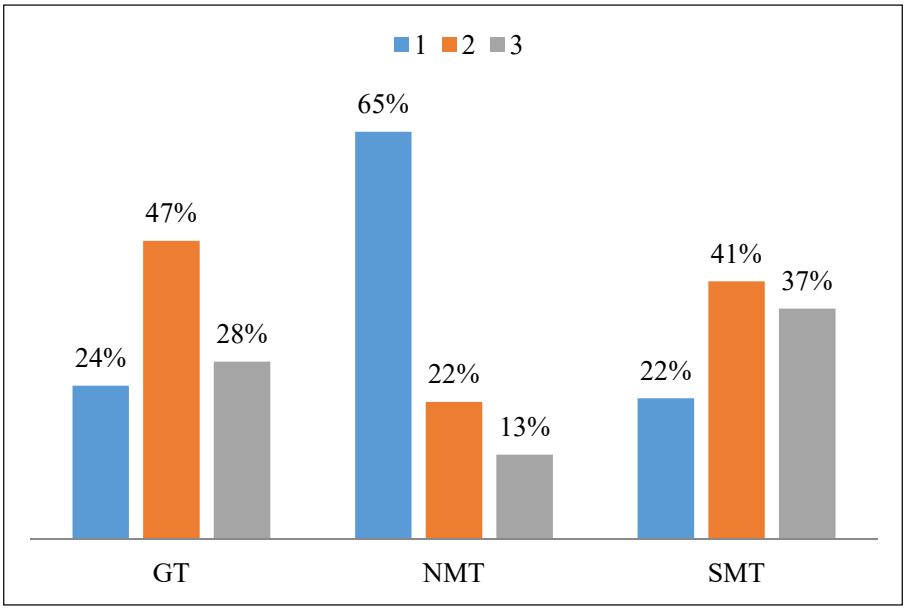
Table 3: Descriptive statistics for fluency

	NMT	GT	SMT
Mean	2.92	1.91	1.91
Median	3	2	2
Mode	3	2	1

Chi-square tests show that there are statistically significant changes in fluency between the three systems: SMT vs. NMT ($X^2(3) = 528.32$, $p < 0.001$), SMT vs. GT

($X^2(3) = 55.38$, $p < 0.001$), NMT vs. GT ($X^2(3) = 541.15$, $p < 0.001$).As far as the ranking task is concerned, the NMT system is selected as the best-rated translation engine in 39.7% of cases, whilst GT is selected in 31% of segments and the SMT system in 29.3%. When looking at the segments selected in the first place, one can see that 24% is linked to GT, 22% to the SMT system and 65% to the NMT system, as shown in Graph 3.

Graph 3: Ranking results



As regards mean values, the NMT system obtains lower values, which in this case show a better assessment as rank 1 represents the best quality and rank 3, the worst: NMT ($M = 1.49$, $SD = 0.72$), GT ($M = 2.04$, $SD = 0.72$) and SMT ($M = 2.14$, $SD = 0.76$).

Table 4: Descriptive statistics for ranking

	NMT	GT	SMT
Mean	1.49	2.04	2.14
Median	1	2	2
Mode	1	2	2

Chi-square tests show that there are statistically significant changes between the three systems regarding their ranking, as shown in Table 5.

Table 5: Ranking results

	Chi-square	Wilcoxon
GT vs. SMT	$X^2(2) = 1{,}583.98$, $p < 0.001$	$Z = 8{,}850$, $p < 0.05$
GT vs. NMT	$X^2(2) = 353.75$, $p < 0.001$	$Z = 18{,}972$, $p < 0.05$
NMT vs. SMT	$X^2(2) = 714.83$, $p < 0.001$	$Z = 3{,}181$, $p < 0.05$

4.2. Quality at the document level

Only 12 replies were recorded for the GT and NMT systems and 13 replies were recorded for the SMT system. Few qualitative comments were added by participants, whose contribution to this task was considered voluntary due to the constraints imposed by the ethical committee.

As regards fluency, mean values show that the NMT system obtains the best ratings ($M = 5$, $SD = 1.65$) in contrast to GT ($M = 3.33$, $SD = 1.15$) and SMT ($M = 3.31$, $SD = 2.17$). In this instance, only the differences between GT and NMT are statistically significant ($U = 33.00$, $p < 0.05$) and between NMT and SMT ($U = 34.50$, $p < 0.05$). Participants provided some contradictory comments concerning the NMT such as "[a]stonishingly, several segments were translated perfectly (10), but then with other segments [*sic*] content was missing or the sentence was not understandable" (P1) or "[n]ot at all fluent, although a few sentences were okay" (P3). Nevertheless, these comments were more positive than those received for the other systems, where participants indicated that "[e]ven simple German sentence structure was not correct, partly the text contained completely untranslated terms" (P3) or "[t]he source quality is not good" (P6). In any case, none of the mean values is high, 5 being the best of all three.

In terms of adequacy, the NMT system again obtains a higher mean value ($M = 6.92$, $SD = 1.44$) compared to the GT ($M = 4.83$, $SD = 2.21$) or the SMT system ($M = 5.54$, $SD = 2.54$). However, differences are only statistically significant when comparing GT and NMT ($U = 30.50$, $p < 0.05$). Participants indicated that in the NMT system "[t]he programm [*sic*] has problems with longer, convoluted sentences and skips parts" (P16) and in the SMT system "[o]ften words in the source language remained, parts of sentences weren't translated at all" (P1).

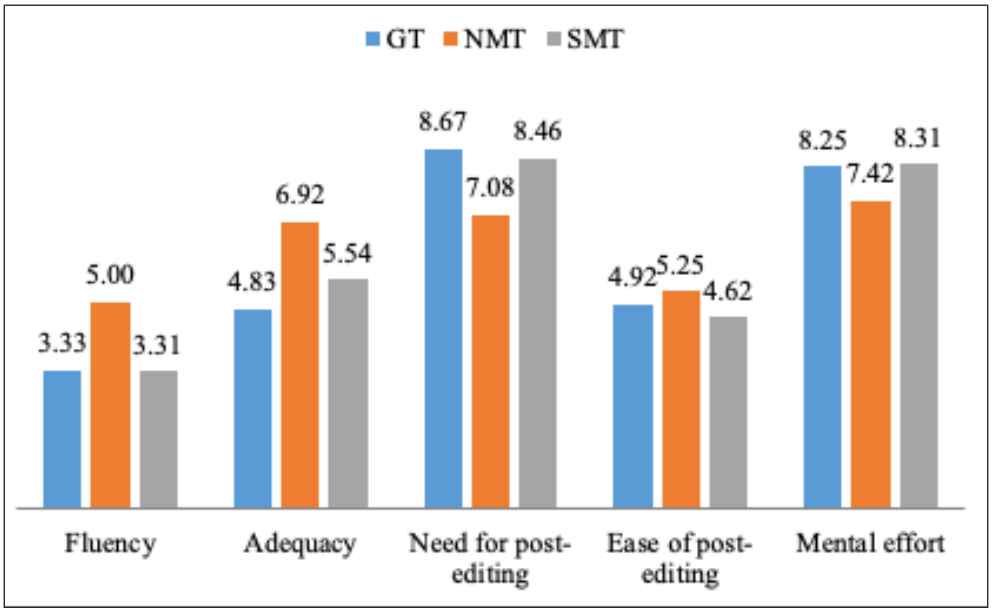
With respect to the need for post-editing, the NMT system (M = 7.08, *SD* = 1.98) is considered to require less post-editing compared to GT (M = 8.67, *SD* = 0.78) and SMT (M = 8.46, *SD* = 1.45). These descriptive results are partially confirmed, since the only statistically significant differences are found between the GT and NMT systems ($U = 33.00$, $p < 0.05$). One participant considered that in the SMT system "[n]early all the segments had to be post-edited" (P20) and another one indicated that in GT "[i]ndividual segments were very well translated, but all in all there was a lot of work" (P1). Regarding the NMT system, comments reflect opposing views: "[i]t required a lot of post editing. The MT text was of low quality and not reliable" (P3) versus "some segments didn't require any editing, others were complicated" (P1).

In relation to the ease of post-editing, mean values for all three systems are quite similar and are not statistically significant: NMT (M = 5.25, *SD* = 2.67), GT (M = 4.92, *SD* = 2.97), and SMT (M = 4.62, *SD* = 3.07). Qualitative data show that the difficulty was the low accuracy of the terminology due to the high degree of specialisation of the texts, which was highly dependent on the knowledge of each participant of the domain. The experimental conditions linked to the chosen tool also had an impact, as participants could not go back to a previous segment and correct it, as mentioned above. In this regard, one participant working with GT indicated that "[i]t was not easy. I had some terminology issues. In a real translation, I would have gone back at the end to change some important terms that I got wrong to make sure they are translated correctly and consistently throughout the file. The result as it is now is awful and would definitely require further editing" (P3). Another participant made the following comment concerning the NMT system: "[t]erminology and context presented the biggest challenge, given that it was impossible to access previously edited segments" (P6). Also, when dealing with the SMT output, similar comments are found: "[i]t was often difficult to understand the whole sentence at once" (P20). It is interesting to notice that one participant indicated that post-editing the NMT system was "[w]ay easier than G3. But the translation still needed some work" (P16). G3 was a GT output.

With regard to mental effort, participants considered that post-editing the NMT output (M = 7.41, *SD* = 2.19) required less effort than post-editing GT (M = 8.25, *SD* = 2.18) and SMT outputs (M = 8.31, *SD* = 1.89), although the differences are not statistically significant. Qualitative data show that the effort was mostly related to the degree of specialisation of the text, as indicated by participant 12 in the following comment in which s/he states that the challenge is "mostly to find out about specialised vocabulary".

Graph 4 presents a summary of mean values for the measures discussed until this point.

Graph 4: Mean values for quality at document level



4.3. Productivity

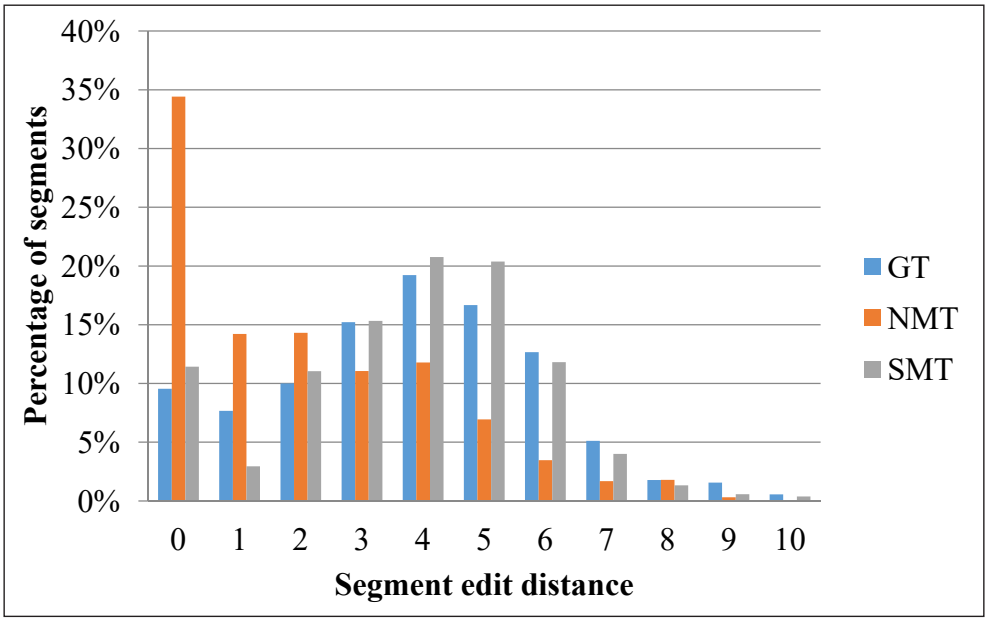
Data from two participants could not be used, as they had not carried out the task following the instructions provided. In terms of post-editing speed, NMT texts were translated at a speed of 1,207 words per hour (*SD* = 630.81), compared to 1,018 words for GT output (*SD* = 500.40) and 996 words per hour for SMT output (*SD* = 483.26). However, a Mann-Whitney U test does not find any statistical differences among the three systems in any case: SMT vs. NMT (U = 160.00 p > 0.05), SMT vs. GT (U = 174.00, p > 0.05), NMT vs. GT (U = 152.00, p > 0.05).

When looking for correlations with the participants' experience as translators, revisers or post-editors –computed through a Spearman's correlation in which the numeric value of the years of experience has been used–, there seems to be no correlation between PE speed and their experience. There is also no correlation between PE speed and PE effort. However, as regards quality assessments at the document level, there seems to be a positive correlation between PE speed and fluency (r_s = -0.59, p < 0.05) and ease of PE (r_s = 0.39, p > 0.05) assessments at the document level in the case of the NMT system, and paradoxically a negative correlation between PE speed and ease of PE (r_s = -0.42, p > 0.05) at the document level in the case of the SMT system.

In terms of post-editing effort, a big difference is observed for 0 edit distance: 34.42% of the segments of the NMT system are included in this range, whilst the

percentage is much lower for the GT (9.56%) and the SMT systems (11.43%). Moreover, the NMT system performs better than GT and SMT in edit distances 1 and 2 (14.21% and 14.32% respectively). Thus, more than 60% of its segments fall in the lowest edit distances, as compared to 27% for GT and 25% for the SMT system. All these values seem to indicate that the NMT is the system that requires less post-editing effort in our experiment. Graph 5 summarises the results.

Graph 5: Edit distances in segments



Mean values on post-editing effort prove the observations in the edit distance distribution: NMT presents a mean of 20.49 (*SD* = 20.82), which is much lower than those of GT (*M* = 37.56, *SD* = 21.51) and SMT (*M* = 37.14, *SD* = 20.21). The differences in terms of PE effort are statistically significant when comparing all systems, as described next: SMT vs. NMT ($U = 266.43$, $p < 0.001$), SMT vs. GT ($U = 479.25$, $p < 0.001$), NMT vs. GT ($U = -212.81$, $p < 0.001$).

Again, when correlating PE effort with the participants' professional experience, there only seems to be a positive correlation between PE effort and their experience as revisers ($r_s = 0.13$, $p < 0.001$). Despite the fact that the correlation is statistically significant, the strength of association is very low.

When correlating PE effort with quality assessments at the document level, statistically significant correlations with a moderate-to-low strength of association are

found for all assessments in the case of the NMT systems. Thus, for higher PE effort values, there are lower values in fluency (r_s = 0.10, p < 0.05), adequacy (r_s = -0.34, p < 0.001) and ease of PE (r_s = -0.21, p < 0.001), but higher values in the need for PE (r_s = 0.25, p < 0.001) and mental effort (r_s = 0.18, p < 0.001). In the case of the SMT and GT systems, a negative correlation exists between PE effort and adequacy (r_s = -0.34, p < 0.001 and r_s = -0.18, p < 0.001 respectively).

4.4. Attitudes

Participants' attitude towards several aspects of MT and PE are not particularly positive, contrary to Cadwell et al.'s (2016) findings in relation to institutional translators from the European Commission's Directorate-General for Translation (DGT). As shown in Table 6, in relation to MT, even though they do not seem to be inclined to use it and think MT texts are of low quality, they paradoxically perceive it to be reasonably useful. As far as PE is concerned, they regard it as a relatively boring task which requires a high cognitive effort, although their interest in PE can be considered fair and their perception of post-edited text quality is actually quite high.

Table 6: Previous attitudes on MT and PE

	Mean	Median	SD
MT quality	2.82	3	1.08
MT usefulness	5	5	2.19
MT use inclination	3.18	2	2.56
Interest in PE	4.55	5	2.42
Boredom associated with PE	4.91	5	2.74
PE cognitive effort	7.55	8	2.07
Quality of PE texts	6.73	7	1.68

A closer analysis of previous attitudes by age (Table 7) shows some interesting results: translators seem to have more positive attitudes towards MT quality and usefulness and to be more inclined to use MT as they grow older. They also seem to be less interested in PE and consider PE more boring as their age increases. However, in terms of their attitude towards PE cognitive effort and the quality of PE texts, there seems to be no age-related pattern.

Table 7: Attitudes according to age

	32-36 years old (4)			37-42 years old (6)			Over 43 years old (11)		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
MT quality	2.00	2	1.41	2.50	2.5	0.58	3.40	3	1.14
MT usefulness	4.00	4	1.41	4.00	4.5	1.41	6.20	7	2.59
MT use inclination	1.00	1	0.00	2.00	2	0.82	5.00	5	2.83
Interest in PE	6.50	6.5	2.12	4.25	4.5	0.96	4.00	2	3.24
Boredom associated with PE	3.00	3	2.83	5.25	5	2.06	5.40	3	3.36
PE cognitive effort	7.50	7.5	3.54	7.25	7	2.22	7.80	8	1.92
Quality of PE texts	6.00	6	1.41	7.00	7.5	1.41	6.80	6	2.17

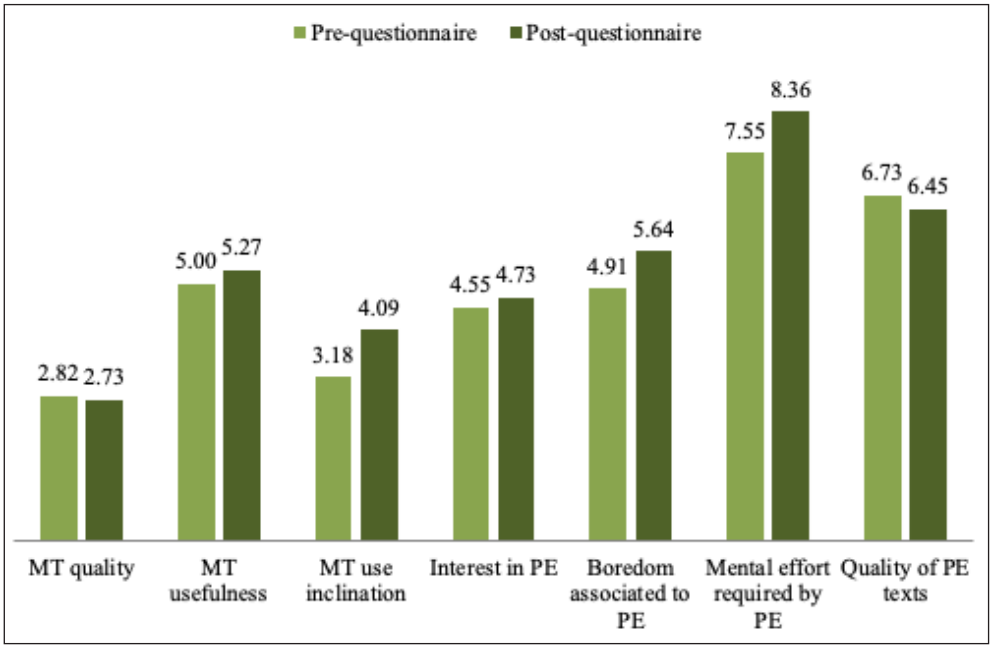
Previous attitudes seem to be related to translators' experience. Although correlations found between different aspects are not statistically significant, the strength of association is moderate. Thus, there is a positive correlation between the attitude towards MT quality and their experience as translators (r_s = 0.46, p > 0.05), as revisers (r_s = 0.34, p > 0.05) and as post-editors (r_s = 0.41, p > 0.05). However, only the experience as revisers presents a moderately positive correlation with the attitude towards MT usefulness (r_s = 0.46, p > 0.05) and towards the inclination to use MT (r_s = 0.39, p > 0.05). Their interest in PE is linked to their experience as post-editors (r_s = 0.35, p > 0.05), while boredom associated with PE has more to do with the experience as revisers (r_s = 0.36, p > 0.05). Nevertheless, it must be noted that the more experience they have as post-editors, the less boring they seem to consider PE (r_s = -0.76, p < 0.01) and the harder they tend to consider it in terms of PE cognitive effort (r_s = 0.37, p > 0.05). Lastly, their attitude toward the quality of post-edited texts is positively correlated with their experience as translators (r_s = 0.50, p > 0.05).

When correlating previous attitudes with the quality assessments at the document level, there are some statistically significant correlations. Thus, it is found that a positive attitude towards the quality of PE texts is correlated with a more positive assessment of the adequacy of texts (r_s = 0.37, p < 0.05). Also, the higher the PE cognitive effort

is thought to be, the lower the ease of PE (r_s = -0.41, p < 0.05) and the higher the perceived PE mental effort (r_s = 0.39, p < 0.05) assessments are. There is also a positive correlation between the mental effort assessment and both the attitudes towards MT usefulness (r_s = 0.361, p < 0.05) and towards the quality of post-edited texts (r_s = 0.46, p < 0.01).

As regards the change of attitudes, participants who did not fill in the questionnaire both before and after the task were not taken into account. A total of 11 replies were collected, which are summarised in Graph 6, where the changes in mean values from the pre-questionnaire to the post-questionnaire are presented.

Graph 6: Changes in attitude: mean values



In most aspects assessed there is a negative change in participants' attitudes. A positive change can only be found when assessing the usefulness of machine translation for translators (from M = 5.00 *SD* = 2.19 to M = 5.27, *SD* = 2.00), the inclination to use machine translated texts as texts as a starting point (from M = 3.18, *SD* = 2.56 to M = 4.09, *SD* = 2.30) and the interest in post-editing (from M = 4.55, *SD* = 2.42 to M = 4.73, *SD* = 2.41). It is worth highlighting that most aspects are assessed with very low rates: the quality of machine translation, which is rated with a 2.82 (*SD* = 1.08), drops to 2.72 (*SD* = 1.35). Post-editing is considered to be more boring (from M = 4.91, *SD*

= 2.74 to *M* = 5.64, *SD* = 3.41) and requiring a higher degree of mental effort (from *M* = 7.55, *SD* = 2.07 to *M* = 8.36, *SD* = 1.69) after carrying out the tasks. Similarly, the quality that professional translators assign to post-edited texts decreases slightly from 6.73 (*SD* = 1.68) to 6.45 (*SD* = 1.63). Statistically significant differences are only found for the mental effort ($Z = -1.84$, $p < 0.05$).

Changes in attitude towards any of the aspects do not appear to be related to any age group in particular. They also do not seem to be related to the years of experience translators have in the translation field or with the fact they have experience revising third-party texts. Previous experience in PE seems to be the only aspect having a statistically significant impact on the positive change in attitude as far as the inclination to use MT is concerned ($Z = -2.03$, $p < 0.05$).

5. Conclusions and summary of results

Results of the comparison of three machine translation systems in the machine-tool domain for the Spanish-German language pair show that the NMT system is ranked highest on all assessed aspects, while GT ranks second in 6 out of the 10 assessed aspects and SMT ranks second in just 4 out of the 10. Table 8 shows all the elements assessed. It indicates that the MT system performs better on a 1 to 3 ranking column and then whether statistically significant differences were found when comparing systems. For attitudes, the symbols used indicate whether a positive or a negative change was found before and after the test, and whether it was statistically significant.

Table 8: Summary of results

MTOOL	**Ranking**			**Significant differences**		
Segment-level Quality	**GT**	**NMT**	**SMT**	**GT vs. NMT**	**GT vs. SMT**	**NMT vs. SMT**
Adequacy	3	1	2	✔	X	✔
Fluency	2	1	3	✔	X	✔
Ranking	2	1	3	✔	✔	✔
Text-level Quality						
Fluency	2	1	3	✔	X	✔
Adequacy	3	1	2	✔	X	X
Need for PE	3	1	2	✔	X	X
Ease of PE	2	1	3	X	X	X
PE mental effort	2	1	3	X	X	X
Productivity						
PE speed	2	1	3	X	X	X
PE effort	3	1	2	✔	✔	✔
Attitude	**Change**					
MT quality	-			X		
MT usefulness	+			X		
MT use inclination	+			X		
Interest in PE	+			X		
PE boredom	+			X		
PE cognitive effort	+			✔		
Quality of PE texts	-			X		

When analysing both automated metrics and human results, it must be pointed out that there is an almost perfect match between the results obtained automatically

and those obtained from the human assessments. NMT is the system which is unanimously awarded the best results. GT and SMT, however, vie for the last position: while GT ranks second in 6 out of the 10 human assessments, it ranks third in the three automated metrics.

Table 9 shows the existing correlations between BLEU and different human assessments. Although the results might not be statistically significant, the strength of the association is what is relevant here (between -1 and +1, indicating negative or positive associations respectively):

Table 9: Correlations between BLEU and human assessments

Adequacy and BLEU	r_s = -0.09, $p > 0.05$
Fluency and BLEU	r_s = 0.48, $p < 0.001$
Ranking and BLEU	r_s = -0.30, $p < 0.001$
PE speed and BLEU	r_s = 0.05, $p > 0.05$
PE effort and BLEU	r_s = -0.50, $p < 0.001$

Leaving aside adequacy and PE speed, where the correlations with BLEU are very low, significant positive correlations are found between fluency and BLEU, and negative correlations are found between the other aspects, so that the greater the fluency, the lower the ranking (hence, a better result) and the lower the PE effort, the higher the BLEU metric, which highlights the existing correlations between PE effort and BLEU.

In view of these findings, we may conclude that the NMT system works much better than both the GT and the SMT systems in a highly technical, specialised domain such as that of machine-tools despite the low amount of both in-domain (25,256) and out-of-domain (1,784,385) training data, which is in line with the findings of other researchers such as Castilho et al. (2017b) in the educational domain, of Wu et al. (2016) for Wikipedia segments, of Bentivogli et al. (2016) for transcribed speeches and of Klubička et al. (2017) in the news field.

As regards subjective opinions and attitudes, some interesting results have been obtained that leave the door open for further research: participants indicate that the quality is highly variable depending on the segments, ranging from perfect translations to unacceptable ones, which shows the potential for automatic quality assessment prior to post-editing to reduce effort and negative assessments. However, their general

attitude towards machine translation and post-editing is not a positive one and does not improve much after the experience. Two aspects that may have had an effect are the experimental design, which did not allow them to go back into the text and make corrections, and the fact that they were dealing with texts of varied quality. Our research has also shown some correlations (or lack of correlations) by age and years of experience in different fields (i.e. translation, revision and post-editing), a field worth exploring in future research in which the assessment of machine translation will hopefully not just rely on automated metrics but also on human factors. In this regard, it is worth the potential in future research of other methodological tools such as focus groups or interviews with different user profiles to obtain more qualitative data on professional users attitudes which can be triangulated with quantitative measures. It also remains to be seen how future training in the area of post-editing will impact on professional attitudes towards this task.

Acknowledgements

This work was partially funded by the Spanish Ministry of Economy and Competitiveness via the AdapTA project (RTC-2015-3627-7). We would like to thank MondragonLingua Translation & Communication as the coordinator of the project and the translators who participated in the experiments. Anna Matamala is a member of Transmedia Catalonia, a research group funded by the Catalan government under SGR call (2017SGR113).

6. References

Axelrod, A., He, X., & Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, 355-362.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *Proceedings of the International Conference on Learning Representations.* CoRR abs/1409.0473.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 29, 65-72.

Bentivogli, L., Bisazza, A., Cettolo, M. & Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 257–267.

Bisazza, A., Ruiz, N., & Federico, M. (2011). Fill-up versus interpolation methods for phrase-based SMT adaptation. *International Workshop on Spoken Language Translation (IWSLT 2011)*, 136-143.

Bryman, A. (2012). *Social Research Methods*. Oxford: Oxford University Press.

Cadwell, P., Castilho, S., O'Brien, S., & Mitchell, L. (2016). Human factors in machine translation and post-editing among institutional translators. *Translation Spaces*, 5, 222-243.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., & Specia, L. (2012). Findings of the 2012 Workshop on Statistical Machine Translation. *Proceedings of the 7th Workshop on Statistical Machine Translation*, 10-51.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017a). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108(1), 109-120.

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Sosoni, V., Georgakopoulou, P., Lohar, P., Way, A., Barone, A.V.M., & Gialama, M. (2017b). A comparative quality evaluation of PBSMT and NMT using professional translators. *Proceedings of MT Summit XVI*, 1, 116-131.

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder–Decoder approaches. *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 103-111.

Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., Akhanov, E., Brunelle, P., Coquard, A., Deng, Y., Enoue, S., Geiss, C., Johanson, J., Khalsa, A., Khiari, R., Ko, B., Kobus, C., Lorieux, J., Martins, L., Nguyen, D., Priori, A., Riccardi, T., Segal, N., Servan, C., Tiquet, C., Wang, B., Yang, J., Zhang, D., Zhou, J., & Zoldan, P. (2016). *Systran's pure neural machine translation systems*. CoRR abs/1610.05540.pdf.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6, 241–252.

Eetemadi, S., Lewis, W., Toutanova, K., & Radha, H. (2015). Survey of data-selection methods in statistical machine translation. *Machine Translation*, 29, 189-223.

Etchegoyhen, T., Fernández-Torné, A., Azpeitia, A., Martínez García, E., & Matamala, A. (2018). Evaluating Domain Adaptation in Machine Translation Across Scenarios. *Proceedings of the Eleventh International Conference on Language Resources Evaluation (LREC 2018)*, 6-15.

Foster, G., & Kuhn, R. (2007). Mixture-model adaptation for SMT. *Proceedings of the Second Workshop on Statistical Machine Translation (StatMT '07)*, 128-135.

Freitag, M., & Al-Onaizan, Y. (2016). *Fast Domain Adaptation for Neural Machine Translation.* CoRR abs/1612.06897.pdf.

Gascó, G., Rocha, M. A., Sanchis-Trilles, G., Andrés-Ferrer, J., & Casacuberta, F. (2012). Does more data always yield better translations? *Proceedings of the 13th European Chapter of the Association for Computational Linguistics*, 152-161.

Görög, A. (2014). Quantifying and benchmarking quality: the TAUS Dynamic Quality Framework. *Tradumàtica*, 12, 443-454.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT '11)*, 187-197.

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A.M. (2017). Opennmt: Open-source toolkit for neural machine translation. *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*, 67-72.

Klubička, F., Toral, A., & Sánchez-Cartagena, V.M. (2017). Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108, 121-132.

Koehn, P., & Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. *Proceedings of the Workshop on Statistical Machine Translation*, 102-121.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. et al. (2007). Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th Annual Meeting of the ACL*, 177-180.

Koponen, M. (2010). Assessing machine translation quality with error analysis. *Electronic proceedings of the KäTu symposium on translation and interpreting studies*, 4, 1-12.

Luong, M.T., & Manning, C.D. (2015). Stanford neural machine translation systems for spoken language domains. *Proceedings of the International Workshop on Spoken Language Translation*, 76-79.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, 50–60.

Moran, J., Saam, C., & Lewis, D. (2014). Towards desktop-based CAT tool instrumentation. *Proceedings of the Third Workshop on Post-Editing Technology and Practice (WPTP-3)*, 99-112.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 311-318.

Schober, P., Boer, C., & Schwarte, L. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, 126 (5), 1763-1768.

Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 539-549

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of Association for Machine Translation in the Americas*, 223-231.

TAUS. (nd.). *Taus dynamic quality framework: Getting started*. Technical report.

Valli, P. (2015). The TAUS Quality Dashboard. *Proceedings of the 37th Conference Translating and the Computer*, 127–136.

Van Peer, W., Hakemulder, F., & Zyngier, S. (2012). *Scientific methods for the humanities*. Amsterdam: John Benjamins Publishing Company.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1, 80–83.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., & Dean, J. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. CoRR abs/1609.08144.