Machine Translation Quality: A comparative evaluation of SMT, NMT and tailored-NMT outputs

Maria Stasimioti
Vilelmini Sosoni
Department of Foreign Languages,
Translation and Interpreting
Ionian University
Corfu, Greece
stasimioti@ionio.gr
sosoni@ionio.gr

Despoina Mouratidis
Katia Kermanidis
Department of Informatics
Ionian University
Corfu, Greece
c12mour@ionio.gr
kerman@ionio.gr

Abstract

The present study aims to compare three systems: a generic statistical machine translation, a generic neural machine translation and a tailored-NMT system focusing on the English to Greek language pair. The comparison is carried out following a mixed-methods approach, i.e. automatic metrics, as well as side-by-side ranking, adequacy and fluency rating, measurement of actual post editing effort and human error analysis performed by 16 postgraduate Translation students. The findings reveal a higher score for both the generic NMT and the tailored-NMT outputs as regards automatic metrics and human evaluation metrics, with the tailored-NMT output faring even better than the generic NMT output.

1 Introduction

Latest technological advances in machine translation (MT) have led to a wider availability of MT systems for various language pairs and neural machine translation (NMT) has been widely hailed as a significant development in the improvement of the quality of MT, given that NMT models have been proven to consistently outperform statistical machine translation (SMT) models in shared tasks, as well as in various project outcomes (Toral and Sánchez-Cartagena, 2017; Castilho et al., 2017a, 2017b, 2018; Klubička et al., 2017, 2018; Popović, 2017, 2018).

MT has been moved "from the peripheries of the translation field closer to the centre" (Koponen, 2016a, p. 131) and has been integrated in the translation workflow, by using machine translated text as a raw translation to be further post-edited by a translator (Lommel and DePalma, 2016; Koponen, 2016b).

The differences between various MT systems, as regards the quality of their output and the types of errors included therein, have been reported by several recent studies. Some (Bahdanau et al., 2015; Jean et al., 2015; Junczys-Dowmunt, 2016; Dowling et al., 2018) relied on automatic evaluation metrics (AEMs) like BLEU (Papineni et al., 2002) and HTER (Snover et al., 2006); others used human evaluations of the MT output quality, employing adequacy and fluency ratings (Bentivogli et al., 2016), manual error analyses (Klubička et al., 2017, 2018; Popović, 2018) or a combination of methods (Burchardt et al., 2017; Castilho et al., 2017a, 2017b, 2018; Toral and Sánchez-Cartagena, 2017; Shterionov et al., 2018; Koponen et al., 2019; Jia et al., 2019; Stasimioti and Sosoni, 2019).

Drawing on these studies, the present study aims to compare three systems: a generic SMT, a generic NMT and a tailored-NMT system, namely a factored or custom-trained NMT system, focusing on the English to Greek language pair. The comparison is performed following a mixed-methods approach, i.e. automatic metrics, as well as side-by-side ranking, adequacy and fluency rating, measurement of actual post-editing (PE) effort and human error analysis. To the best of our knowledge there are no studies to date for the English to Greek language pair which compare generic and custom-trained MT systems, while there are only a few related studies to date comparing SMT and NMT systems (Castilho et al., 2017b; Stasimioti and Sosoni, 2019)

2 Methodology

A mixed-methods approach was adopted in the present study with a view to producing reliable results. AEMs and human evaluation metrics, including eye-tracking and keystroke logging data for measuring the effort expended by translators while carrying out full PE of the MT output

generated by three different systems (Google Translate SMT system, Google Translate NMT system, tailored-NMT system), side-by-side ranking of the MT outputs, adequacy and fluency rating and human error classification were used to evaluate the quality of the MT output of these three MT systems and investigate their differences.

A series of experiments were carried out during the 2018-2019 Spring Semester at the Department of Foreign Languages, Translation and Interpreting of the Ionian University. Twenty Greek students enrolled on the MA in the Science of Translation initially participated in this study. However, only sixteen completed all tasks, since the participation in the tasks was optional. All participants signed a consent form, while all stored data were fully anonymised in accordance with Greek Law 2472/97 (as amended by Laws 3783/2009, 3917/2011 and 4070/2012).

2.1 Participants and training

As can be seen in Table 1, all participants were female, the majority belonged to the 18-24 and 25-34 age groups, they all had an undergraduate degree either in Translation or in a related field, while only five of the participants had professional experience in translation. In addition, none of the participants had experience in PE.

Gender	Female	16
Age distribution	18-24	10
	25-34	5
	45-54	1
Education level	Undergraduate degree holder	13
	Postgraduate degree holder	2
	PhD holder	1
Degree type	Translation	9
	Other	7
Experience in Translation	Yes	4
	No	12
Experience in PE	Yes	0
	No	16

Table 1. Participants' gender, age distribution, education level, degree type and experience in translation and PE

PE training was a prerequisite for participating in this study. For that reason, specific training was offered in the context of the compulsory module "Translation Tools" and aimed to introduce students to MT and PE as well as to the recent developments in the respective fields. Upon completion of the training, students were expected, among others, to be able to (i) use MT during the pre-translation process, (ii) evaluate MT output using both automatic and human evaluation metrics and (iii) post-edit MT output according to the expected level of quality (full/light PE).

To that end, the topics covered included, among others, the theory and history of MT and PE, the basic principles of MT technology, analysis of the dominant systems in the market, the importance of controlled language and pre-editing for MT, quality metrics and evaluation of MT output, PE levels of quality, PE effort and productivity (temporal, technical and cognitive effort), MT output error identification, MT engine implementation in the translation workflow and post-editor profile and associated skills (O'Brien, 2002; Depraetere, 2010; Doherty et al., 2012; Doherty and Kenny, 2014; Kenny and Doherty, 2014; Koponen, 2015; Guerberof and Moorkens, 2019).

2.2 Source Texts

The source texts (STs) used in this study were 4 short (~140 words) semi-specialised texts about the 2019 EU elections selected from the British daily newspaper *The Guardian*. They all had comparable Lexile® scores (between 1200L and 1300L), i.e. they were suitable for 11th/12th graders (see Table 2). The Lexile Analyzer[1] was used as it relies on an algorithm to evaluate the reading demand – or text complexity – of books, articles and other materials.

	Text 1	Text 2	Text 3	Text 4
Lexile® Measure	1200L–1300L	1200L–1300L	1200L–1300L	1200L–1300L
Number of sentences	7	6	8	8
Mean sentence length	20.86	23.50	20.00	19.71
Word count	146	141	140	138

Table 2. Lexile® scores for the STs used in the study

[1] https://la-tools.lexile.com/free-analyze/

2.3 MT systems

As already mentioned, for the present study we used three different MT systems: the SMT system developed by Google (Google Translate SMT system - GSMT), the NMT system developed by Google (Google Translate NMT system - GNMT) and a tailored-NMT system. The first two are generic MT systems, i.e. general purpose systems, trained with huge amounts of data from various subject areas and thus suitable to translate texts in all subject areas or domains. Google Translate, in particular, is the best known MT service, which can be used either free of charge as a standalone tool (translate.google.com) or for a small fee via an API for translating large amounts of text or for using it within a CAT tool. The third system is a custom-trained system developed by Kanavos and Nadalis (2019) with the Open NMT toolkit (Klein et al., 2017) and trained with publicly available parallel corpora, including a parallel corpus compiled from the RAPID multilingual parallel corpus compiled from all press releases of the Press Release Database of European Commission released between 1975 and end of 2016[2] as well as a parallel corpus of English and South-East European Languages which is based on the content published on the SETimes.com[3] news portal. Although generic MT systems provide "reasonable quality" (Vasiljevs et al., 2016: 134) for many language pairs (Aiken, 2019), they do not perform particularly well for domain and user-specific texts and are much less effective than custom-trained MT systems, which in most cases produce better results (Ping, 2009).

2.4 Evaluation

In order to evaluate the MT output generated by each MT system we used both AEMs and human evaluation metrics.

2.4.1 Automatic Evaluation Metrics (AEMs)

The AEMs used in this study were BLEU, METEOR, WER and TER. BLEU measures the similarity between the MT output and a reference translation, METEOR (Lavie and Agarwal, 2007) is based on the weighted harmonic mean of unigram precision and recall, while WER (Zechner and Waibel, 2000) and TER are based on Levenshtein distance and calculate the number of edits required to make an MT output match the reference translation. It should be noted that we used two (2) reference translations by professional translators, since the use of a single human-translated reference tends to introduce bias (Popovic et al., 2016).

2.4.2 Human Evaluation

As pointed out, human evaluation included eye-tracking and keystroke logging data for measuring the effort expended by translators while carrying out full PE of each MT output, side-by-side ranking of the MT outputs, adequacy and fluency rating and error classification. As regards the PE, the participants were asked to perform full PE of the MT output (either the output from the generic SMT system, the generic NMT system or the tailored-NMT system) of four semi-specialised texts, which were presented to them in a random order. All participants were asked to rank and rate for adequacy and fluency all the segments from each MT output for all four texts (87 segments in total) and perform a classification of all the errors found therein. Each participant performed the tasks in one go, starting from the PE and moving on to the ranking, rating and error analysis tasks. The questionnaires were filled right after the completion of the PE tasks.

2.4.2.1 Measurement of PE effort (temporal, technical and cognitive)

According to Krings (2001), there are three categories of PE effort: (i) temporal effort, (ii) technical effort and (ii) cognitive effort (Krings, 2001, p.179). Cognitive effort is directly related to temporal effort and technical effort.

For the aims of this study, the participants were asked to carry out PE tasks while the temporal effort (total task time), the technical effort (keystrokes: insertions and deletions) and the cognitive effort (number of fixations, mean fixation duration and total gaze time) expended were registered using a Tobii X2-60 remote eye-tracker and the Translog-II software (Carl, 2012). The effectiveness of using eye-tracking as an MT evaluation technique has been proven by previous studies (Doherty et al., 2010). Although using eye-tracking involves humans, much of the subjectivity involved in human evaluation of MT quality is removed as the processes that eye-tracking measures are largely unconscious (Doherty et al., 2010).

Prior to the execution of the tasks, a group meeting was organised during which the participants were informed about the nature of the experiments, the task requirements and the general as well as task-specific guidelines they

[2] http://europa.eu/rapid/

[3] http://www.setimes.com

had to follow. In particular, the participants were asked to carry out full PE of the MT output generated by the aforementioned three MT systems, according to the task-specific guidelines, i.e retain as much raw MT translation/output as possible, transfer the message accurately, fix any omissions and/or additions (at the level of sentence, phrase or word), correct mistranslations, correct morphological errors, correct misspellings and typos, fix incorrect punctuation if it interferes with the intended message, correct erroneous terminology, fix inconsistent use of terms and do not introduce stylistic changes. The task began with a warm-up PE task which aimed to familiarise each participant with the procedure; the data from the warm-up task were not included in the ensuing analysis and discussion. The actual experimental task involved the full PE of the MT output of four semi-specialised texts by each one of the participants. The texts for full PE were presented to the participants in a random order. During the experiment, the ST was displayed in the Translog-II software at the top half of the screen and the MT output at the bottom half. The participants were asked to carry out the tasks at the speed at which they would normally work in their everyday work as translators; therefore, no time constraint was imposed. In addition, they worked directly on the MT output.

2.4.2.2 Side-by-side ranking

After the eye-tracking experiments the participants were given a side-by-side task for each text (Text 1, Text 2, Text 3 and Text 4) and were asked to read the Greek translations of each English source segment carefully and rank them in order from best to worst. The SMT, NMT and tailored-NMT outputs were presented to participants using Google Forms in a random order.

2.4.2.3 Adequacy and fluency

Following the ranking task, the participants were asked to rate each segment from each MT output for all four texts (87 segments in total) for adequacy and fluency (defined as the extent to which a target segment is correct in the target language and reflects the meaning of the source segment) on a five-point Likert scale for each segment.

In particular, the translators were asked to rate adequacy in response to the question "Is the MEANING of the English sentence kept in the translation?". A five-point Likert scale was used, where 1 is "Not at all", 2 is "Barely", 3 is "Partly", 4 is "Mostly" and 5 is "Fully". Similarly, the translators were asked to rate fluency in response to the question "Considering only GRAMMAR and SPELLING, the translated sentence is:". Like in the case of adequacy, a five-point Likert scale was used where 1 is "Very poor", 2 is "Poor", 3 is "Fair", 4 is "Good" and 5 is "Excellent".

2.4.2.4 Error classification

The last task for the participants was an error classification task. The error typology used in this study was suggested by Stasimioti and Sosoni (2019) and was a combination of the subset of the Dynamic Quality Framework (DQF) and Multidimensional Quality Metrics (MQM) harmonized error typology suitable for MT analysis as suggested by Lommel and Melby (2018) and the MQM error typology which was widely used in previous studies mainly due to the flexibility of the error types and their granularity (Klubička et al., 2017; 2018; Carl and Báez, 2019).

In particular the participants were asked to classify the errors of each segment in two main error categories and their subcategories; adequacy errors: addition, omission, mistranslation, untranslated text, terminology error and fluency errors: error in grammar, error in punctuation, error in style, spelling error and typo.

3 Findings and discussion

3.1 Automatic Evaluation Metrics (AEMs)

Table 3 shows the scores of the AEMs we used per system.

	SMT	NMT	tailored-NMT
BLEU	0.34	0.39	**0.46**
METEOR	0.48	0.52	**0.56**
WER	0.50	0.49	**0.43**
TER	0.52	0.51	**0.39**

Table 3. Average of AEMs per system

The tailored-NMT system outperformed both the SMT and the NMT systems. In further detail, it is observed that the tailored-NMT output shared more common words with the reference translations (higher BLEU score) than did the SMT and NMT outputs. In addition, the higher METEOR score observed at both segment and system levels in the tailored-NMT output showed that there are significant matches between words and phrases in the tailored-NMT output and the reference translations. As regards TER and WER,

the majority of edits observed were substitutions and deletions. The tailored-NMT system achieved the lowest score between the systems. Given that TER and WER are edit-distance metrics, a lower score indicates better performance. As far as the SMT and NMT systems are concerned, the latter performed better in all cases achieving higher BLEU and METEOR scores and lower TER and WER scores.

3.2 Human evaluation

3.2.1 Measurement of PE effort (temporal, technical and cognitive)

Temporal effort

As far as the temporal effort is concerned, we measured the average time (in minutes) the participants needed to post-edit each MT output. As it emerges from Figure 1, the MT output generated by the tailored-NMT system required less time for full PE (M = 8.73, SD = 3.16) compared to the MT outputs generated by the NMT system (M = 9.85, SD = 3.55) and the SMT system (M = 12.76, SD = 5.11). A one-way ANOVA was conducted to compare the effect of the MT output on temporal effort (task duration) when post-editing the SMT output, the NMT output and the tailored-NMT output. There was a significant effect of the MT output on temporal effort for these three conditions $F(2,45) = 4.28$, $p = 0.019$. Post hoc comparisons indicated that mean task duration when post-editing the SMT output was significantly different, i.e. higher, than mean task duration when post-editing the tailored-NMT output. However, mean task duration when post-editing the NMT output did not significantly differ from mean task duration when post-editing the SMT output and the tailored-NMT output.

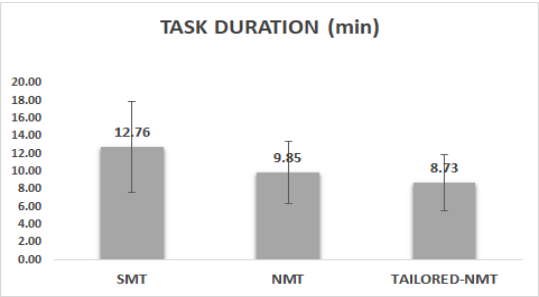

Figure 1. Temporal effort: Mean and standard deviation of task duration per system

Technical effort

Technical effort is generally measured by the number of keystrokes, which can be distinguished into insertions and deletions. As it emerges from Figure 2 and similarly to temporal effort, the participants performed fewer keystrokes when post-editing the tailored-NMT output (M = 228, SD = 114) compared to the keystrokes performed when post-editing the NMT output (M = 362, SD = 116) and the SMT output (M = 520, SD = 208). A one-way ANOVA yielded a statistically significant difference $F(2,45) = 14.72$, $p < 0.05$ for the average number of keystrokes (insertions and deletions), as well as for the insertions $F(2,45) = 14.18$, $p < 0.05$ and deletions $F(2,45) = 14.12$, $p <$ 005 separately. Post hoc comparisons indicated that the average number of keystrokes performed when post-editing the SMT output was significantly different, i.e. higher, than the average number of keystrokes performed when post-editing the NMT output and the tailored-NMT output. In addition, the average number of keystrokes performed when post-editing the tailored-NMT output was significantly different, i.e lower, than the average number of keystrokes performed when post-editing the NMT output. The same applies for the insertions and deletions separately.

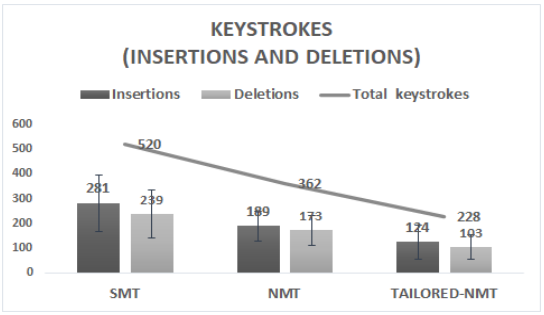

Figure 2. Technical effort: Mean and standard deviation of keystrokes per system

Cognitive effort

Pause duration and pause density (Lacruz and Shreve, 2014; Daems et al., 2017; Koponen et al., 2019; Jia et al., 2019), fixation count, fixation duration and gaze time (Mesa-Lao, 2014; Moorkens et al., 2015) have been used in previous studies as indicators of cognitive effort. In our study we measured the average fixation count, the mean fixation duration (in milliseconds) as well as the average total gaze time (in minutes), i.e. the sum of all fixation durations, on both areas of the screen (ST at the top half of the screen and MT output at the bottom half of the screen) in order to compare the cognitive effort expended by the translators when post-editing each MT output. As far as the average fixation count is concerned (see Figure 3), this was higher when post-editing the SMT output (M = 1460, SD = 547) than the NMT output (M = 1154, SD = 421) and the tailored-NMT output (M = 1089, SD = 375). Apart from

the higher average fixation count, the SMT output also triggered longer gaze time (M = 7.97, SD = 2.90) than the NMT (M = 6.32, SD = 2.31) and the tailored-NMT output (M = 5.59, SD = 1.86) (see Figure 4). The mean fixation duration was exactly the same when post-editing the SMT and the NMT output (M = 329, SD = 34 and M = 329, SD = 28 respectively) and slightly lower when post-editing the tailored-NMT output (M = 311, SD = 33) (see Figure 5). A one-way ANOVA yielded a statistically significant difference $F(2,45) = 4.11$, $p = 0.023$ for the total gaze time, but not for the number of fixations $F(2,45) = 3.05$, $p = 0.057$ or the mean fixation duration $F(2,45) = 1.63$, $p = 0.206$. Post hoc comparisons indicated that total gaze time when post-editing the SMT output was significantly different, i.e. longer, than total gaze time only when post-editing the tailored-NMT output. In addition, the average fixation count when post-editing the SMT output was significantly different, i.e. higher, than the average fixation count only when post-editing the tailored-NMT output.

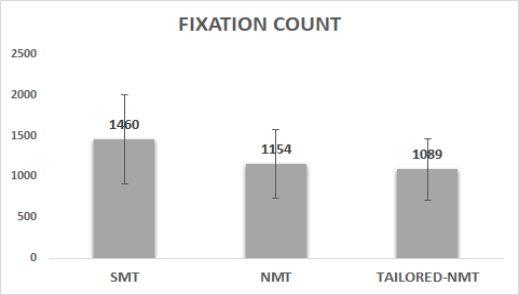

Figure 3. Cognitive effort: Mean and standard deviation of fixation count per system

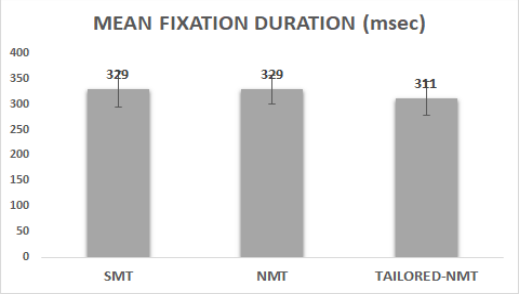

Figure 4. Cognitive effort: Mean and standard deviation of mean fixation duration per system

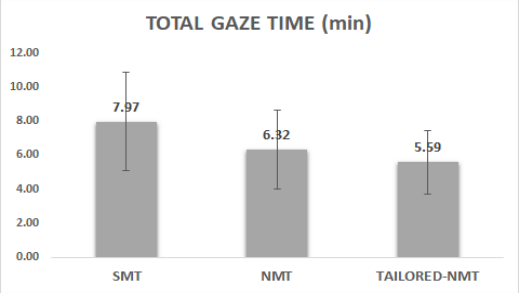

Figure 5. Cognitive effort: Mean and standard deviation of total gaze time per system

3.2.2 Side-by-side ranking

As it emerges from Figure 6, the tailored-NMT output was ranked as the best by 37% of the participants, compared to 34% for the SMT output and 29% for the NMT output. The SMT output was ranked as the worst by almost half (45%) of the participants, while the NMT output was ranked second by 53% of the participants. It is observed that quite a high percentage, namely 38% of the participants, ranked tailored-NMT output as the worst. This may be explained by the higher number of omissions and punctuation errors found in the output as can be seen in the error classification in 3.2.4. To assess the agreement between the annotators we computed Fleiss' kappa coefficient (Fleiss, 1971). Inter-annotator agreement shows fair agreement among the annotators (κ = 0.40).

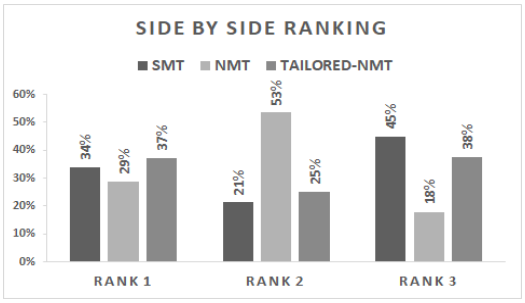

Figure 6. Average percentage of ranking per system

3.2.3 Adequacy and fluency

As it emerges from Figure 7, the tailored-NMT output was rated higher for both adequacy and fluency followed by the NMT output. In particular, both the tailored-NMT and the NMT outputs were deemed to be good by the translators/annotators, both as regards the grammaticality and the conveyance of meaning, while the SMT output was deemed to be fair in both respects. Inter-annotator agreement shows fair agreement among the annotators for fluency (κ = 0.29) and slight agreement for adequacy (κ = 0.10).

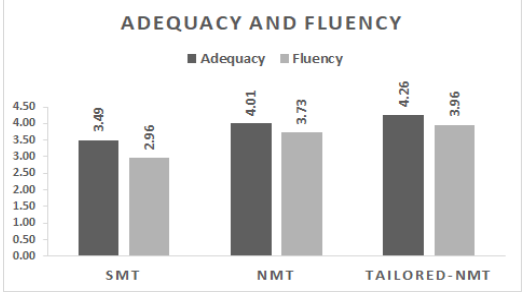

Figure 7. Weighted average of adequacy and fluency rating per system

3.2.4 Error classification

As far as the number of errors is concerned (see Figure 8), the tailored-NMT output contains the lowest number of errors overall, while the SMT output contains the highest number of errors overall.

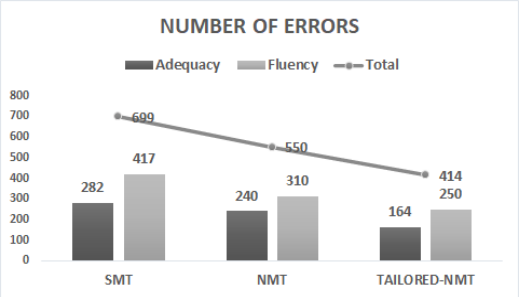

Figure 8. Total number of errors per system

As far as the types of errors are concerned (see Figures 8 and 9), we observed that all MT outputs contain more errors at the level of fluency than at the level of adequacy.

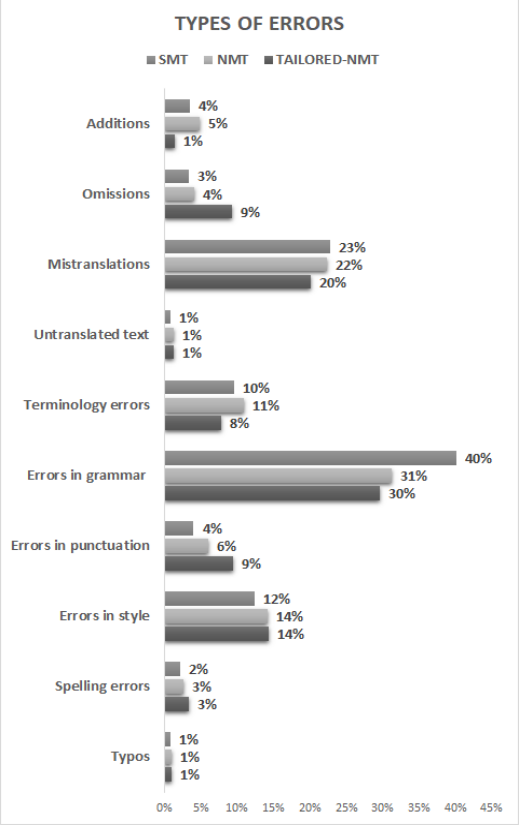

Figure 9. Average percentage of error types per system

As regards the category of fluency, the SMT output contains significantly more grammatical errors than the NMT and tailored-NMT outputs. Another interesting finding as regards fluency involves the category of punctuation. The tailored-NMT output contains almost 60% more punctuation errors than the SMT output and almost 30% more than the NMT output. This difference is due to the fact that the em dashes found in the STs are omitted in the tailored-NMT output in all cases. As far as the category of adequacy is concerned, the tailored-NMT output contains slightly fewer mistranslations, terminological errors and additions but more omissions than the NMT and SMT outputs. Inter-annotator agreement shows fair agreement among the annotators (κ = 0.22).

4 Discussion and conclusion

This paper reports on a comparative evaluation of generic SMT, generic NMT and tailored-NMT outputs for the English to Greek language pair using AEMs and human evaluation metrics, including eye-tracking and keystroke logging data, side-by-side ranking of the MT outputs, adequacy and fluency rating and error classification. As regards the differences between the SMT and NMT outputs, this study shows that the NMT systems produce translations of higher quality and thus corroborates the findings of previous studies on various language pairs (Toral and Sánchez-Cartagena, 2017; Klubička et al., 2017, 2018; Jia et al., 2019; Koponen et al., 2019), including the English-Greek language pair (Castilho et al., 2017a, 2017b; Stasimioti and Sosoni, 2019). In particular, the analysis reveals a higher score for both the generic NMT and the tailored-NMT outputs as regards automatic metrics and human evaluation metrics, with the tailored-NMT output faring even better than the generic NMT output. In addition, the tailored-NMT output was ranked as the best and was rated higher for both adequacy and fluency, a fact which explains the reduced temporal, technical and cognitive effort expended during its PE.

The decrease in PE effort can also be explained by the lowest number of errors found in the tailored-NMT output. Another interesting finding is that all the MT outputs contain more errors at the level of fluency than at the level of adequacy with the most typical fluency errors being grammatical errors. Both NMT outputs contain fewer grammatical errors than the SMT output, confirming thus the findings of previous studies for improved quality of the NMT systems at the level of fluency - not only for the English-Greek language pair (Castilho et al., 2017a, 2017b; Stasimioti and Sosoni, 2019) but also for other language-pairs, such as English-Czech, English-German, English- Finnish, English-Romanian,

English-Russian, English-Croatian and English-Chinese (Toral and Sánchez-Cartagena, 2017; Klubička et al., 2017, 2018; Jia et al., 2019). However, no difference between the generic NMT and the tailored-NMT outputs was reported. Another interesting finding involves the category of punctuation where the tailored-NMT output fares worse than both the generic NMT and the generic SMT output. Finally, in terms of adequacy, the tailored-NMT output fares better with slightly fewer mistranslations, terminological errors and additions than the generic NMT and the generic SMT outputs, although it includes more omissions than the former.

The findings point to the fact that there are limits to generic MT models, as they are not tuned to provide translations that are unique to a specific genre and thus business or industry. Although the development of a tailored-NMT system can be particularly compute intensive, and therefore too expensive and time-consuming – especially in cases where there are not enough parallel data to train a new good quality and appropriately adapted system – the higher quality and the reduced cognitive, technical and temporal effort suggest that it is worth exploring further.

Acknowledgement We would like to thank the HUBIC Lab (Raptis and Giagkou, 2016) for providing the Tobii X2-60 remote eye-tracker for the purposes of this study.

References

Aiken, Milam. 2019. An Updated Evaluation of Google Translate Accuracy. *Studies in Linguistics and Literature*. 3:253.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations*. San Diego, CA, USA.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In EMNLP 2016 *Proceedings of the 2016 conference on empirical methods in natural language processing*. Austin, Texas, 1-5 November 2016. Association for Computational Linguistics, 257–267.

Burchardt, Aljoscha, Vivien Macketanz, Jon Degdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. 2017. A linguistic evaluation of rule-based, phrase-based, and neural MT engines. *Prague Bulletin of Mathematical Linguistics*, 108: 159–170.

Carl, Michael. 2012. Translog – II: A program for recording user activity data for empirical reading and writing research. In *Proceedings of the 8th international conference on language resources and evaluation*. Istanbul, Turkey, 21-27 May 2012. European Language Resources Association (ELRA), 4108–4112.

Carl, Michael, and María Cristina Toledo Báez. 2019. Machine Translation Errors and the Translation Process: A Study across Different Languages. *Journal of Specialised Translation*, 31:107-132.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017a. Is neural machine translation the new state of the art? *Prague Bulletin of Mathematical Linguistics*, 108:109–120.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Miceli Barone, and Maria Gialama. 2017b. A Comparative quality evaluation of PBSMT and NMT using professional translators. In *Proceedings of Machine Translation Summit XVI*, vol.1: Research Track. Nagoya, Japan, 18-22 September 2017. 16th Machine Translation Summit, 116–131.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Yota Georgakopoulou, Andy Way, Antonio Miceli Barone, and Maria Gialama. 2018. Evaluating MT for massive open online courses: A multifaceted comparison between PBSMT and NMT systems. *Machine Translation*, 32(3):255–278.

Daems, Joke, Sonia Vandepitte, Robert J. Hartsuiker, and Macken, Lieve. 2017. Identifying the Machine Translation Error Types with the Greatest Impact on Post-editing Effort. *Frontiers in Psychology.* 8: 1282.

Depraetere, Ilse. 2010. What counts as useful advice in a university post-editing training context? Report on a case study. In François Yvon and Viggo Hansen (eds). 2010. *EAMT 2010: Proceedings of the 14th annual conference of the European association for machine translation*, 27-28 May 2010 Saint-Raphaël Congrès, Saint-Raphaël, France, European Association for Machine Translation.

Doherty, Stephen, Sharon O' Brien, and Michael Carl. 2010. Eye tracking as an Automatic MT Evaluation Technique. *Machine Translation*. 24:1-13.

Doherty, Stephen, Dorothy Kenny, and Andy Way. 2012. Taking statistical machine translation to the student translator. In AMTA-2012 *The Tenth Biennial Conference of the Association for Machine Translation in the Americas*. San Diego, USA, 28 October-1 November 2012. AMTA, n.p.

Doherty, Stephen, and Dorothy Kenny. 2014. The design and evaluation of a statistical machine translation syllabus for translation students. *The Interpreter and Translator Trainer*, 8(2):295–315.

Dowling, Meghan, Teresa Lynn, Alberto Poncelas and Andy Way. 2018. SMT versus NMT: preliminary comparisons for Irish. In *AMTA 2018 Workshop: LoResMT 2018*, 17-21 Mar 2018, Boston, MA. USA.

Guerberof, Anna and Joss Moorkens. 2019. Machine translation and post-editing training as part of a master's programme. *Jostrans: The Journal of Specialised Translation*, 31:217–238.

Fleiss, Joseph. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378-382.

Jia, Yanfang, Michael Carl, and Xiangling Wang. 2019. Post-editing neural machine translation versus phrase-based machine translation for English–Chinese. *Machine Translation* (2019).

Jean, Sébastien, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal Neural Machine Translation Systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, 134–140.

Junczys-Dowmunt, Marcin, Dwojak Tomasz and Hoang Hieu. 2016. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. In *Proceedings of the International Workshop on Spoken Language Translation 2016.* 1 edn, vol. 1, 4, Japan.

Kanavos, Panos, and Costas Nadalis. 2019. MT Systems and MT Training. Presentation at the *Crash Course in Machine Translation* organized by Dimetra Academy. 26-28 June 2019, Athens, Greece.

Kenny, Dorothy, and Stephen Doherty. 2014. Statistical machine translation in the translation curriculum: Overcoming obstacles and empowering translators. *The Interpreter and Translator Trainer*, 8(2), 276–294.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart and Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Vancouver, Canada, 67-72.

Klubička, Filip, Antonio Toral, and Victor M. Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *Prague Bulletin of Mathematical Linguistics,* 108:121–132.

Klubička, Filip, Antonio Toral, and Victor M. Sánchez-Cartagena. 2018. Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian. *Machine Translation*, 32(3): 195–215.

Koponen, Maarit. 2015. How to teach machine translation post-editing? Experiences from a post-editing course. In *Proceedings of 4th Workshop on Post-Editing Technology and Practice* (WPTP4). Miami, United States, 30 October-3 November 2015. AMTA, 2–15.

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

Koponen, Maarit. 2016b. *Machine translation post-editing and effort: Empirical Studies on the post-editing effort*. PhD. Helsinki: University of Helsinki.

Koponen, Maarit, Leena Salmi, and Markku Nikulin. 2019. A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Machine Translation* (2019).

Krings, Hans. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes*. Kent: Kent State University Press.

Lacruz, Isabel, and Gregory M. Shreve. 2014. Pauses and cognitive effort in post-editing. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia. (eds.) *Post-editing of machine translation*. Newcastle: Cambridge Scholars Publishing.

Lavie, Alon, and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second ACL Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 228-231

Lommel, Arle, and Donald A. DePalma. 2016. *Europe's leading role in Machine Translation: How Europe is driving the shift to MT. Technical report*. Boston: Common Sense Advisory.

Lommel, A. and Melby, A.K. (2018). MQM-DQF: A good marriage (Translation quality for the 21st century). Tutorial at the *13th Conference of The Association for Machine Translation in the Americas*. Boston, USA, 17 March 2018. AMTA. https://www.aclweb.org/anthology/W18-1925

Mesa-Lao, Bartolome. 2014. Gaze behaviour on source texts: An exploratory study comparing translation and post-editing. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia, (eds.) *Post-editing of machine translation*. Newcastle: Cambridge Scholars Publishing.

Moorkens, Joss, Sharon O'Brien., Igor A.L. Silva, Norma Fonseca, and Fabio Alves. 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation* 29(3): 267–284.

O'Brien, Sharon. 2002. Teaching post-editing: A proposal for course content. In EAMT 2002 *Proceedings of the 6th annual conference of the European association for machine translation*. Manchester, UK, 14-15 May 2002. European Association for Machine Translation, 99–106.

Papineni, Kishore, Roukos Salim, Ward Todd, and Zhu Wei-Jing. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, Philadelphia, 311-318.

Ping, Ke. 2009. Machine translation. In Mona Baker, and Gabriela Saldanha (Eds.), *Routledge encyclopedia of translation studies* 2: 162 168. London: Routledge.

Popovic, Maja, Mihael Arčan and Arle Lommel. 2016. Potential and Limits of Using Post-edits as Reference Translations for MT Evaluation. *Baltic J. Modern Computing*, 4(2):218-229.

Popović, Maja. 2017. Comparing language related issues for NMT and PBMT between German and English. *Prague Bulletin of Mathematical Linguistics*, 108(1):209–220.

Popović, Maja. 2018. Language-related issues for NMT and PBMT for English–German and English–Serbian. *Machine Translation*, 32(3):273–252.

Raptis, Spyros, and Maria Giagkou. 2016. From capturing to generating human behavior: closing the interaction loop at the HUBIC Lab. In *Proceedings of the 20th Pan-Hellenic Conference on Informatics (PCI) with International Participation*, Patras, Greece, 10-12 November 2016. ACM Digital Library, International Conference Proceedings Series.

Shterionov, Dimitar, Riccardo Superbo, Pat Nagle, Laura Casanellas, and Tony O'Dowd. 2018.. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation* 32**:**217–235.

Snover, Matthew, Bonnie Dorr, Schwartz Richard, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, The Association for Machine Translation in the Americas, Cambridge, 223-231.

Stasimioti, Maria, and Vilelmini Sosoni. 2019. MT output and post-editing effort: Insights from a comparative analysis of SMT and NMT output for the English to Greek language pair and implications for the training of post-editors. In C. Szabó & R. Besznyák (eds) *Teaching Specialised Translation and Interpreting in a Digital Age - Fit-For-Market Technologies, Schemes and Initiatives*. Wilmington: Vernon Press.

Stymne, Sara. 2011. BLAST: A Tool for Error Analysis of Machine Translation Output. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 19-24 June, Portland, Oregon, USA - System Demonstrations, 56-61.

Toral, Antonio, and Victor M. Sánchez-Cartagena 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th conference of the European Chapter of the Association for Computational Linguistics*: Volume 1, Long Papers. Valencia, Spain, 3-7 April 2017. ACL, 1063–1073.

Vasiļjevs, Andrejs, Jan Hajic, Jochen Hummel, Josef van Genabith, and Rihards Kalniņš. 2016. European Platform for the Multilingual Digital Single Market: Conceptual Proposal. In Inguna Skadiņa and Roberts Rozis (eds.) *Human Language Technologies – The Baltic Perspective*. Amsterdam: IOS Press BV.

Zechner, Klaus and Alex Waibel. 2000. Minimizing word error rate in textual summaries of spoken language. *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.