Preferences of end-users for raw and post-edited NMT in a business environment

Sabrina Girletti, Pierrette Bouillon
FTI/TIM, University of Geneva
sabrina.girletti@unige.ch
pierrette.bouillon@unige.ch

Martina Bellodi, Philipp Ursprung
Swiss Post Ltd
martina.bellodi@post.ch
philipp.ursprung@post.ch

Abstract

This paper presents an evaluation conducted with end-users of translations produced by Swiss Post's in-house Language Service. The aim is to assess whether end-users i) would rate post-edited MT more highly than raw MT; ii) would find that Swiss Post's customized NMT system produces better results than a general-purpose, off-the-shelf NMT engine (DeepL) and, lastly, when aware of translation production metadata, iii) would be willing to pay for post-edited texts. This latter aspect in particular was intended to help determine whether the customers would still value human intervention or whether they would rather accept a lower quality translation and associated risks if this means they can save on costs. Results show that the post-edited texts are preferred by the majority of the participants, even when production metadata are revealed. The in-house customized engine seems to produce better results than DeepL, since the end-users choose the raw output from our system more often than from DeepL.

1 Introduction

The in-house Language Service at Swiss Post translates a wide variety of internal and external texts from and into German, French, Italian and English. After evaluating different machine translation (MT) solutions over the past two years (Bouillon *et al.*, 2018), Swiss Post's Language Service is now ready to introduce neural machine translation (NMT) into its workflow. During the testing phase, a customized system built on OpenNMT-tf (Klein *et al.*, 2017) and trained with more than 2.5 million in-domain sentences[1] was found to produce raw translations that are reusable for post-editing purposes in three language directions. In this first phase, machine translation output quality and fitness for purpose was solely evaluated by in-house professional translators. Nevertheless, the Language Service was also interested in determining its customers' (end-users') opinions on machine translation.

As emerged from internal discussions over the years, the extensive hype around neural machine translation, due to its higher fluency in comparison with previous approaches (Wu *et al.*, 2016; Hassan *et al.*, 2018; Castilho *et al.*, 2017), has led many Swiss Post employees to turn to freely available, generic MT systems to obtain quick raw translations. A crucial point in the use of such systems concerns personal data and information security, as the employees are likely to paste sensitive company information into web-based MT interfaces and are not always aware of the risks of this practice.

Therefore, we decided to carry out a study to assess whether end-users a) would rate post-edited MT more highly than raw MT; b) would find that Swiss Post's customized MT system produces better results than a generic MT engine (DeepL) and, lastly, when aware of some production metadata, such as security and cost, c) would be willing to pay for post-edited texts. This latter aspect could help determine whether the customers would still value human intervention or whether they would rather accept a lower quality and associated risks if this means they can save on costs. Our main hypotheses are that end-users will prefer post-edited versions over the

[1] We used the same training data detailed in (Bouillon *et al.*, 2018)

raw MT output, will find that our customized system produces better results than DeepL and will be willing to pay for a translation revised by a professional translator and produced in a secure environment.

The remainder of this paper is structured as follows. Section 2 provides a brief overview of previous studies involving end-users of machine-translated texts, while Section 3 details the methodology of our study. Findings are presented and discussed in Section 4, before concluding.

2 Previous work

There are relatively few studies available that consider the opinion of end-users of machine-translated texts. Many organizations and language services survey their end-users and customers about the quality of their MT output, but the results are not always shared (Bowker and Ehgoetz, 2007).

In their literature review, Bowker and Ehgoetz (2007) cite the studies conducted by Senez (1998) at the European Commission, by Vasconcellos and Bostad (1992) at the Pan American Health Organization and by Nuutila (1996) at Nokia. All reported that end-users (administrative staff and employees of the company) were mostly satisfied with the raw or post-edited output, especially due to the fact that they could receive texts that would not otherwise have been translated. However, the authors do not provide any detail on the methodology followed.

More recent work on the topic has been carried out by Castilho (2016), Castilho and O'Brien (2017), Screen (2019) and Van Egdom and Pluymaekers (2019). In these studies, end-users evaluated the quality of human translations and MT output with different degrees of post-editing. Nevertheless, the participants were never asked to consider additional information about the translations before assessing the texts.

The idea of sharing production metadata, such as production method, time and cost, with end-users in a recipient MT evaluation was first introduced by Bowker and Ehgoetz (2007) and further investigated in Bowker (2009) and Bowker and Ciro (2015).

Bowker and Ehgoetz (2007) surveyed 31 anglophone professors from the Faculty of Arts at the University of Ottawa to determine whether they would accept English translations of lower quality for internal texts, if these were cheaper and faster to obtain than human translations. The source texts selected were administrative memos in French, used only internally. The participants were presented with three versions of the same text (a human translation, a raw MT produced with the Systran MT system (v4.0) and a lightly post-edited version) and with information about the time and cost required to produce them. Two thirds of the respondents preferred the post-edited version, and one third the human translations, while raw MT did not receive any preference. The authors suggested that the acceptability rate of raw MT could be different among groups of participants that have "less of a vested interest in language per se" (p. 221).

This intuition proved correct in Bowker (2009), who conducted a similar study with members of two Canadian official language minority communities. The findings show that, for gisting purposes, lay users more often select raw or lightly post-edited MT output, whereas, when translation is considered as a means of cultural preservation, full PE and human translation are required. Interestingly, the main group selecting this latter option turned out to be language professionals. Therefore, the author states that "average recipients are more open to the idea of MT than are language professionals" (p. 148).

Adopting the framework of community-based participatory research, Bowker and Ciro (2015) conducted a recipient evaluation with 114 Spanish-speaking newcomers at the public library in Ottawa. The participants had to evaluate four different Spanish translations (raw MT produced by Google Translate, fully or lightly post-edited MT, human translation) of a text extracted from the

library's website in English and choose the version that best met their needs. Immediately afterwards, production metadata were provided and the participants were asked to confirm or change their preference. The results show that on average, in the first part, human translation and full post-editing combined were selected by 66% of the respondents, while the raw MT was the less chosen category. Once production method, time and cost were revealed, the preference given to human translation and full post-editing decreased to 23%, while that for raw MT grew considerably. In the end, when considering time and cost for production, lightly post-edited texts proved to be the most chosen ones.

In our study, we propose a similar methodology to investigate end-users' preferences for raw and post-edited NMT when considering production metadata (detailed in Section 3.5). Furthermore, we indirectly compare the output produced by a general-purpose and a customized NMT system.

3 Methodology

In order to answer our research questions, we carried out two comparative evaluations of raw and post-edited versions of machine-translated texts extracted from Swiss Post's manuals. The MT systems tested were an in-house customized NMT system, namely Swiss Post NMT, abbreviated as SPNMT, and a general-purpose, off-the-shelf NMT system, namely DeepL. The language directions tested were German to French and German to Italian. In this section, we will provide information on the participants, the test data, and detail the methodology of the two evaluations.

3.1 Participants

Participants were recruited through a call for participation on the company's intranet and by means of convenience sampling. They were eligible to take part if they 1) were Swiss Post employees, and 2) had French or Italian as their mother tongue. These participants actively need translation of work instructions and manuals in their daily work and are indirect Swiss Post Language Service customers. Thirty-nine Swiss Post employees volunteered for the study, including 23 native French speakers, 15 native Italian speakers and one native bilingual participant who asked to take part in both languages. Therefore, there were forty participants in total.

In a final demographic questionnaire, the participants stated that their level of comprehension of German was moderate or advanced. 28% of the participants reported never using MT, while for the remaining 72%, a half reported using MT "regularly" in their daily work and a half only "sometimes". Participants who reported using MT commented that they do it mostly for gisting purposes or as an aid in their daily work while writing emails in a foreign language or translating texts (even if they do not work for the company as language experts). The systems mentioned were DeepL and Google Translate. Many participants also commented on the fact that they never reuse raw MT as such, as they are usually not satisfied with the results of these systems.

3.2 Test data

The test set consisted of eight texts randomly extracted from Swiss Post's manuals, which are used only internally at Swiss Post. The intended readership is employees in Swiss Post branches (points of sale). Typically, the manuals describe new products or services that Swiss Post is offering or new processes that are relevant for the point of sale. The texts ranged between 212 and 305 words each, for a total of 2130 words and 180 segments. They had been translated from German into French and Italian using our customized NMT system (Swiss Post NMT, abbreviated as SPNMT) and DeepL. Details on the test set are shown in Table 1 below.

Text	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	**Total**
Words	212	303	228	298	305	296	292	196	2130
Segs	18	25	17	24	27	21	28	20	180

Table 1: Number of source words and segments in the test set.

Two in-house professional translators (one per language pair) working at the Language Service were asked to perform a full post-edit of the raw output produced by the customized system and DeepL. HTER scores (Snover *et al.*, 2006) calculated on the post-edited versions and reported in Table 2 below suggest that the post-editors made fewer corrections to the raw output produced by the customized system than to that of DeepL. As shown in Table 2, BLEU scores (Papineni *et al.*, 2002) computed on the raw MT output were systematically higher for our customized system than for DeepL, suggesting higher similarity between the output of the former and the official human reference. Regarding the type of corrections that the post-editors made to the raw output, some examples of sentences to evaluate are reported in Table 3. In particular, in the first example, we note that the post-editor did not make any modifications to the raw output produced by Swiss Post NMT, since this system suggested a much more accurate translation than the one proposed by DeepL.

	DE–FR		**DE–IT**	
System	SPNMT	DeepL	SPNMT	DeepL
BLEU	41.52	28.36	37.01	23.92
HTER	15.36	26.19	20.97	31.18

Table 2: BLEU and HTER scores per system and language pair.

3.3 Test design

All the participants assessed the raw and PE versions of the eight source texts in the same order. Four texts had been machine-translated using our customized NMT solution, while the other four had been machine-translated using DeepL. We split the participants into two sub-groups per target language. Participants in Group 1 assessed translations produced by Swiss Post NMT in texts from 1 to 4, while texts from 5 to 8 came from DeepL. For participants in Group 2, the order of the systems was reversed, as shown in Table 4. Group 1 for French had 13 subjects, and Group 2 had 11 subjects. For Italian, Group 1 and Group 2 contained 8 subjects each. This design applies to both the first and the second evaluation.

Texts	**Group 1**	**Group 2**
1-4	SPNMT	DeepL
5-8	DeepL	SPNMT

Table 4: Test design.

3.4 First evaluation: preference based on quality

The aim of the first comparative evaluation was to assess whether the end-users found that 1) post-editing could significantly improve raw output and that 2) the customized system could produce a more acceptable output than DeepL.

Source (example 1)
Jedes Team verf¨ugt ¨uber eine Leitung Filiale und eine Stellvertretung.
Raw Swiss Post NMT
Chaque équipe dispose d'une direction de filiale et d'une suppléance.
PE Swiss Post NMT
Chaque équipe dispose d'une direction de filiale et d'une suppléance.
Raw DeepL
Chaque équipe est composée d'un *directeur d'agence* et d'un *directeur adjoint.*
PE DeepL
Chaque équipe est composée d'un *responsable de filiale* et d'un *suppléant.*
Source (example 2)
Die Leitung Filiale ist Ansprechperson gegen Aussen und trâgt die Gesamtverantwortung.
Raw Swiss Post NMT
La direzione filiale è la persona di contatto verso l'esterno e assume la responsabilitàgenerale.
PE Swiss Post NMT
Il responsabile Filiale è la persona di contatto verso l'esterno e assume la responsabilitàgenerale.
Raw DeepL
La direzione della filiale è il referente esterno e ha la responsabilitàgenerale.
PE DeepL
Il responsabile Filiale è il referente esterno e ha la responsabilitàgenerale.

Table 3: Examples of sentences to evaluate from the first evaluation. Each participant received the source and a couplet of translations (raw and PE version) from the same system, either Swiss Post NMT or DeepL.

In this evaluation, the participants compared the translations sentence by sentence (180 segments per participant). They did not know which was the raw or the PE version, nor if the translations originated from Swiss Post NMT or from DeepL. The participants performed the evaluation in a questionnaire on the Limesurvey platform. The instructions, sent via email, together with the link to the evaluation, required the participants to read the source sentence in German, then the two translation proposals, and decide which translation they preferred. Available options were:

- *I prefer translation A*
- *I prefer translation B*
- *Both are acceptable*
- *Neither of them*

The questionnaires were sent to the participants on 28th January 2019, together with the instructions in their own mother tongue, namely French or Italian.

3.5 Second evaluation: preference based on quality and production metadata

The aim of this second evaluation was to ascertain whether, when being aware of the parameters for the production of the two versions, the end-users would still prefer the same version as the first evaluation or would change their choices. In particular, we were interested in finding out whether Swiss Post employees were willing to pay for post-edited translations.

This time, the participants were presented with the entire texts (source and target translations), instead of just sentences (although in a coherent order). Differences between the raw and the PE versions were highlighted in blue. In the same interface, we revealed information about:

- the *results of the first evaluation*, namely which version received more preferences on average and was therefore considered *clearly* or *slighly better* than the other[1]
- the *production method*, namely whether the text was "revised by a professional translator" or " non-revised");
- the *server* which hosted the MT system used, namely if this was in-house ("secure") or external ("not secure", as the data would exit the company premises) and, finally;
- the *price* for the raw and the post-edited version. These provisional figures were established by Language Service management for the purposes of this study and calculated per source word. The raw versions produced by Swiss Post NMT and DeepL would cost approximately 86% and 94% less than their PE versions, respectively. Therefore, if a PE version would cost CHF 1, the corresponding raw version produced by the customized system would cost CHF 0.14, while the raw version produced by DeepL would cost CHF 0.06.

The criteria mentioned were all deemed important for the end-user. In particular, the results of the first part relate to the quality of the output, the production method reveals the human intervention on a machine-translated text, the information about the server location raises users' awareness regarding the issue of data confidentiality, and the cost is often one of the main concerns in a business scenario.

Considering these metadata, participants were asked to select the text for which they would pay.

Available options were:

- *I would pay for translation A*
- *I would pay for translation B*
- *I would not pay for either of them*

When choosing this latter option, participants were asked to comment on the reason why, in their opinion, neither of the two translations was worth its price. This evaluation was carried out on the Limesurvey platform three weeks after the first one. The participants received via email detailed instructions on how to perform the task.

4 Results and discussion

In this section, we will present and discuss the results obtained in the two comparative evaluations described in Section 3.

[1] We decided to present the results of the first part in terms of slight or clear preference for one of the two translations. In case of a tie (``both are acceptable"), we would have looked at the second most chosen category.

4.1 Results of the first evaluation

The results of the first evaluation, represented in Figure 1 below, show an overall higher preference for the post-edited version over the raw one, for both MT systems in question. However, while for DeepL there is a clear preference for the post-edited output over the raw one (56% and 13%, respectively), results for our customized system show that, most of the time, the raw and post-edited segments are considered equally acceptable (44% of the judgements), with PE being the second most chosen category (34%) and raw output collecting around 10% of the preferences. These tendencies are the same for both language pairs, as detailed in Figures 2 and 3. These results confirm the HTER and BLEU scores mentioned in Section 3.2: the post-editors had to make fewer corrections to the raw output produced by the customized system, as this was already more similar to the human reference translation, compared to DeepL.

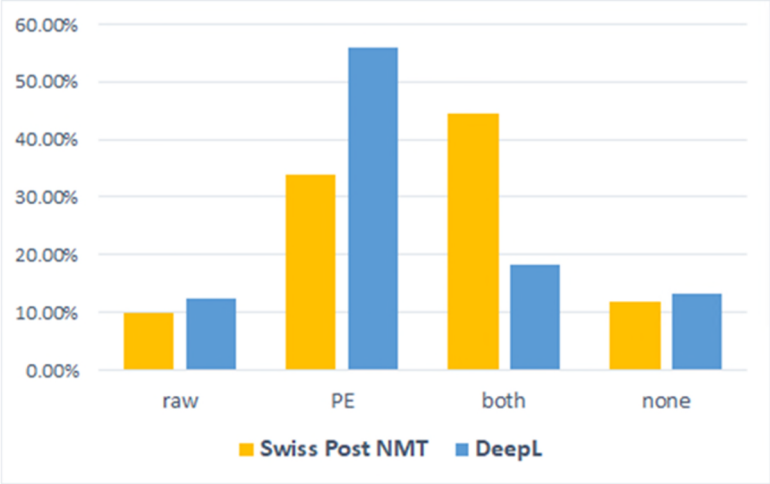

Figure 1: Overall preferences for Swiss Post NMT and DeepL in the first evaluation.

An Inter-Rater Reliability (IRR) analysis was performed using Light's Kappa coefficient (Light, 1971) to assess consistency among nominal ratings provided by the participants in each group, per target language. The results, interpreted according to Landis and Koch (1977), show moderate agreement for the French-speaking participants (k = 0.44 and k = 0.49 for group 1 and group 2, respectively) and fair agreement for the Italian-speaking participants k = 0.33 and k = 0.38 for group 1 and group 2, respectively).

These results are perhaps not surprising, since it is well known that customized systems outperform generic systems (Koehn and Knowles, 2017; Khayrallah *et al.*2018), but they shed light on interesting aspects. For instance, the overall amount of sentences in the categories "raw" and "none" is quite high and deserves attention. As the goal of post-editing is to improve the raw output while correcting any errors, it could be argued that, overall, in more than 11% of the cases, the post-editing made the translation worse (raw better) and in 12% of the cases, PE was not effective enough (neither of the translations is preferred). When analyzing these sentences in greater depth, we see that they contained punctuation issues or small typos, as shown in the examples below:

SOURCE - Hinweis: TWINT-Voucher d¨urfen nicht [...]
RAW - Remarque: les bons TWINT ne doivent pas [...] PE - Remarque: les bons TWINT ne doivent pas [...]

SOURCE - Mit der Bildung von Teamorganisationen [...]
RAW - Con la creazione di organizzazioni del team [...] PE -
con la formazione di organizzazioni [...]

Furthermore, the fact that evaluators rated one sentence at a time could have led to misinterpretations, as in the example below, where the correction rightly made by the post-editors to the paragraph title was not perceived as an improvement:

SOURCE - TWINT-Voucher stornieren
RAW - Annuler *le* bon TWINT
PE - Annuler *un* bon TWINT

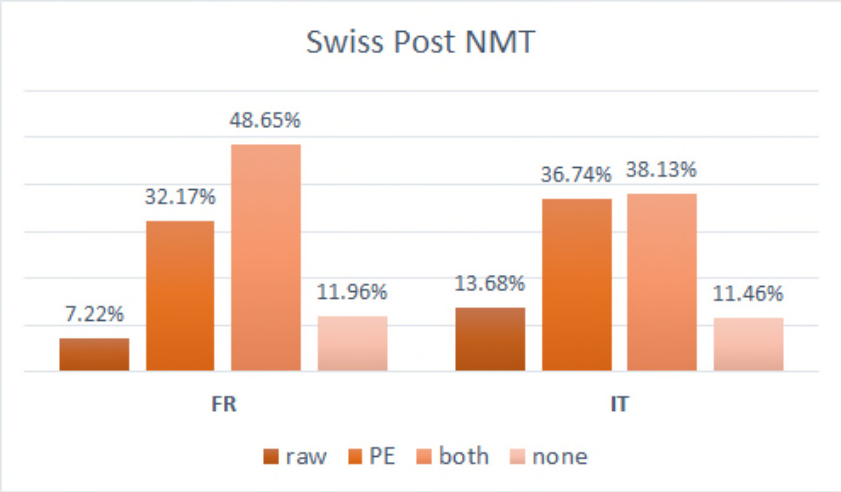

Figure 2: Detail of preferences for Swiss Post NMT in the first evaluation, per target language (German as source language).

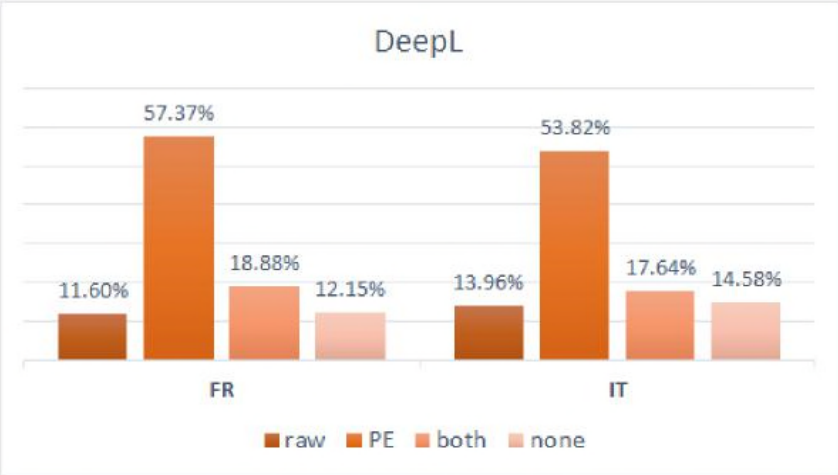

Figure 3: Detail of preferences for DeepL in the first evaluation, per target language (German as source language).

4.2 Results of the second evaluation

In the second test, the participants were asked to make a choice between the raw and post-edited version, or to state that they would not choose either of them. Once metadata related to production method, the results of the first test, server and cost are revealed, the post-edited version is still the most chosen version for both Swiss Post NMT and DeepL (79% and 88%, respectively). Detailed results for language pairs and systems are shown in Figures 5 and 6 below.

Interestingly, we notice an overall increase in the percentage of preferences to the raw output of the customized system (15%), a decrease for the same category in the case of DeepL (around 7%), as well as an overall decrease for the "none" category. The difference observed could be due to the fact that this time the participants made a choice regarding the whole text and not only individual segments. However, while in the first evaluation we wanted to gather more fine-grained results about the quality of the output, the purpose of the second evaluation was to focus on the impact of metadata and to let the participants make a judgement on the translated text in its entirety. For the same reason, in comparison to the first evaluation, we have eliminated one category ("both are acceptable") and forced the participants to express their choice to pay for one of the two versions or neither of them.

Since we did not separate the metadata, we cannot establish if the choice is due to a specific criterion or to a combination of some of them. Nevertheless, as stated previously, all the production metadata can be considered relevant to the end-users. The price, in particular, seems to play a prominent role. In fact, when looking at the reasons why the participants selected the option "I would not pay for either of them", the comments reveal that the evaluators often recognize that the post-edited version is slightly better than the raw one, but some improvements could still be made. For the customized system in particular, since many raw segments were acceptable and therefore have not been modified, some participants clearly stated that the price difference between the raw and post-edited versions was too large and did not reflect the difference in quality.

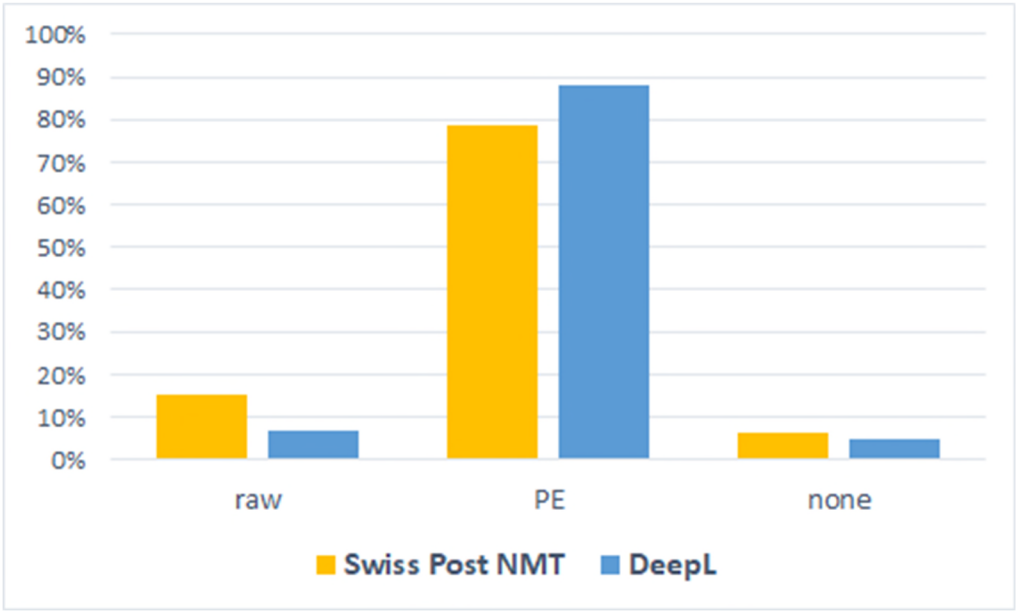

Figure 4: Overall preferences for Swiss Post NMT and DeepL in the second evaluation.

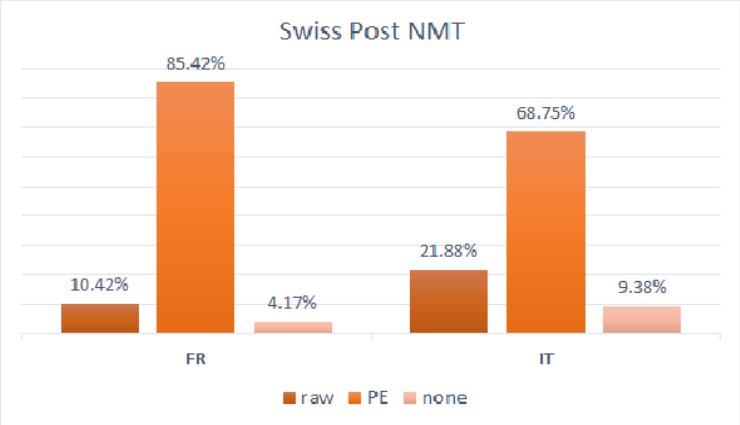

Figure 5: Detail of preferences for Swiss Post NMT in the second evaluation, per target language (German as source language).

Out of a total of 320 judgements collected[2] in the second evaluation, in 70 cases (around 22% of the times) the participants change their preferences between the first and the second evaluation. Details of these changes are shown in Figure 7 below.

For texts translated with DeepL, there is a higher number of changes (7) from the categories "PE" or "raw" to the category "none", compared to the changes in the same direction for texts translated with Swiss Post NMT (5). Conversely, changes from the categories "raw" or "none" to the category "PE" occur more often for Swiss Post NMT (18) than for DeepL (8).

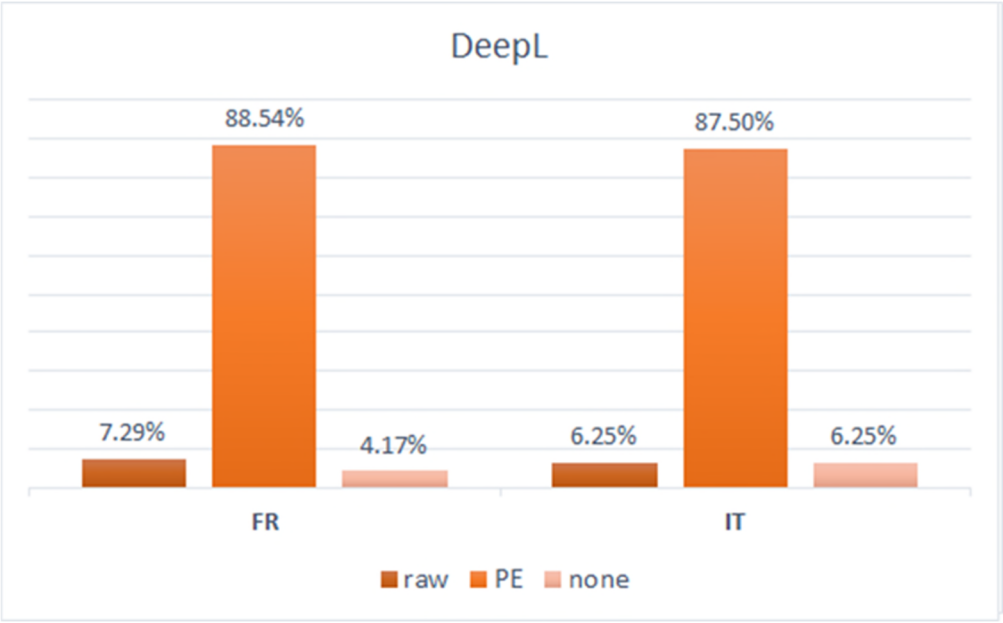

Figure 6: Detail of preferences for DeepL in the second evaluation, per target language (German as source language).

The highest number of changes overall (23) can be found for texts translated with Swiss Post NMT and confirms the results mentioned in Section 4.2: when assessing texts coming from the in-house

[2] 40 participants, 8 texts to assess.

customized system, participants are more likely to accept raw output instead of the post-edited output or none. Only one participant changed his preferences systematically, always choosing the raw version over the post-edited one, and regardless of the NMT system, clearly showing that the price had a major impact on his choices.

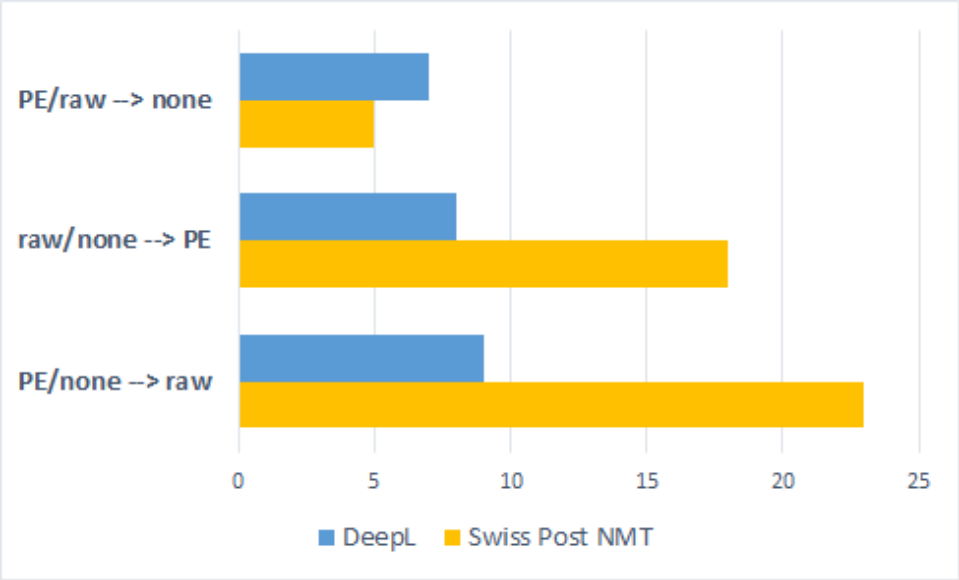

Figure 7: Number of times a participant changed his preference between the first and the second evaluation, per system.

5 Conclusion and future work

In this study, we have reported on the results of two comparative evaluations carried out to investigate Swiss Post employees' preferences regarding raw and post-edited MT from an in-house, customized engine (Swiss Post NMT) and a general-purpose, off-the-shelf MT system (DeepL). We first asked the end-users to express their preference based on the quality of the output, then verified whether they would change these preferences when aware of some production metadata for the two translations.

Despite a rather small sample of participants (n=40), our findings show that post-edited texts are preferred by the majority of evaluators, even when production metadata such as server security and price are revealed. The end-users do perceive a difference in quality between the output of our in-house customized system and that of DeepL, since in the first evaluation the post-edited and raw versions produced by the former are judged equally acceptable more often. In the second evaluation, participants change their preferences more often for Swiss Post NMT, choosing to pay for the raw version over the post-edited one. This could be due to the fact that, in the case of the customized system, the price difference between the two versions is not reflected in the quality difference, as emerged from some participants' comments. Although only a minority of participants expressed this view, the question of how to establish the price of post-edited machine translation in a Language Service remains one that should be carefully considered.

This initial study was subject to a number of limitations: first, we tested only one text type, namely Swiss Post's internal manuals, and two language pairs. It is likely that, for other text types, the rate of preferred raw output could be different. Second, we provided only a full post-edited version of the raw output, however, it would be interesting to repeat the study with different degrees of post-editing and more than one post-edited version. Furthermore, it was not possible to evaluate the

impact of a specific production parameter, as these were all presented together, and we did not include other metadata that could be useful for the end-user's decision, such as the production time. Nonetheless, to the best of our knowledge, this is the first study conducted in a real business setting and investigating end-users' preferences for raw or post-edited NMT output after disclosing production metadata. The results of the study indicate that the quality of the output of a customized trained NMT system is higher compared to a generic non-trained system in the given business scenario. Also, it seems that end-users have more confidence in a customized trained system that addresses data security and are therefore more willing to accept raw MT output compared to the untrained system.

Acknowledgments

We would like to thank all Swiss Post employees who took part in the evaluation, as well as the in-house language experts who provided the translations for the study and reviewed the final version of the paper. We also thank the anonymous reviewers for their insightful comments that helped improve this work.

References

Bouillon, Pierrette, Estrella, Paula, Girletti, Sabrina, Mutal, Jonathan, Bellodi, Martina, Bircher, Beatrice. 2018. Integrating MT at Swiss Post's Language Service: preliminary results. In: *Proceedings of the 21st Annual Conference of the European Association for Machine Translation,* 281–286.

Bowker, Lynne. 2009. Can machine translation meet the needs of official language minority communities in Canada? A recipient evaluation. In: *Looking For Methodological Issues in Translation Studies* edited by Sonia Vandepitte, Antwerpen: Artesis Hogeschool, 123–155.

Bowker, Lynne and Buitrago Ciro, Jairo. 2015. Investigating the usefulness of machine translation for newcomers at the public library. *Translation and Interpreting Studies,* 10(2), 165–186.

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

Castilho, Sheila. 2016. Measuring acceptability of machine translated enterprise content. PhD thesis. Dublin City University.

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

Castilho, Sheila and O'Brien, Sharon. 2017. Acceptability of machine-translated content: A multi-language evaluation by translators and end-users. *Linguistica Antverpiensia,* New Series: Themes in Translation Studies, 16, 120-–136.

Hassan, Hany, Aue, Anthony, Chen, Chang, Chowdhary, Vishal, Clark, Jonathan, Federmann, Christian, Huang, Xuedong, Junczys-Dowmunt, Marcin, Lewis, William, Li, Mu, Liu, Shujie, Liu, Tie-Yan, Luo, Renqian, Menezes, Arul, Qin, Tao, Seide, Frank, Tan, Xu, Tian, Fei, Wu, Lijun, Wu, Shuangzhi, Xia, Yingce, Zhang, Dongdong, Zhang, Zhirui, and Zhou, Ming. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *Computing Research Repository,* arXiv:1803.05567.

Khayrallah, Huda, Thompson, Brian, Duh, Kevin and Koehn, Philipp. 2018. Regularized Training Objective for Continued Training for Domain Adaptation in Neural Machine Translation. *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation,* ACL, Association for Computational Linguistics. Vancouver, Canada. pp. 28–39.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. arXiv preprintarXiv:1701.02810.

Koehn, Philipp and Knowles, Rebecca. 2017. Six Challenges for Neural Machine Translation. *Proceedings of the 1st Workshop on Neural Machine Translation,* ACL, Association for Computational Linguistics. Melbourne, Australia. pp. 36–44.

Landis, J Richard and Koch, Gary G. 1977. The measurement of observer agreement for categorical data. *Biometrics,* JSTOR, pp. 159–174.

Light, Richard. 1971. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological bulletin* vol. 76 nr. 5, American Psychological Association.

Nuutila, Pertti. 1996. Roughlate service for in-house customers. *Papers from the Aslib conference,* held on 14 and 15 November 1996. London: Aslib, 1996.

Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing. 2002. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th ACL,* 311–318, ACL.

Screen, Benjamin. 2019. What effect does post-editing have on the translation product from an end-user's perspective? *JoSTrans, Journal of Specialised Translation* 31, 133–157.

Senez, Dorothy. 1998. Post-Editing Service for Machine Translation Users at the European Commission. *Translating and the Computer 20,* Proceedings from Aslib conference, 12 and 13 November 1998, London.

Snover, Matthew, Dorr, Bonnie, Schwartz, Richard, Micciulla, Linnea, and Makhoul, John. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of Association for Machine Translation in the Americas,* Cambridge, MA:AMTA, 223–231.

Van Egdom, Gys-Walt and Pluymaekers, Mark. 2019. Why go the extra mile? How different degrees of post-editing affect perceptions of texts, senders and products among end-users. *JoSTrans, Journal of Specialised Translation* 31, 158–176.

Vasconcellos, Muriel and Bostad, Dale A. 1992. Machine translation in a high-volume translation environment. In John Newton (ed.) *Computers in translation: a practical appraisal,* London, Routledge, 58–77.

Wu, Yonghui, Schuster, Mike, Chen, Zhifeng, Le, Quoc V., Norouzi, Mohammad, Macherey, Wolfgang, Krikun, Maxim, Cao, Yuan, Gao, Qin, Macherey, Klaus, Klingner, Jeff, Shah, Apurva, Johnson, Melvin, Liu, Xiaobing, Kaiser, Lukasz, Gouws, Stephan, Kato, Yoshikiyo, Kudo, Taku, Kazawa, Hideto, Stevens, Keith, Kurian, George, Patil, Nishant, Wang, Wei, Young, Cliff, Smith, Jason, Riesa, Jason, Rudnick, Rudnick, Vinyals, Oriol, Corrado, Greg, Hughes, Macduff and Dean, Jeffrey. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *Computing Research Repository* arXiv:1609.08144.