Measuring Comprehension and User Perception of Neural Machine Translated Texts: A Pilot Study

Lieve Macken
Ghent University
Groot-Brittanniëlaan 45
9000 Ghent, Belgium
Lieve.Macken@ugent.be

Iris Ghyselen
Ghent University
Groot-Brittanniëlaan 45
9000 Ghent, Belgium
Iris.Ghyselen@ugent.be

Abstract

In this paper we compare the results of reading comprehension tests on both human translated and raw (unedited) machine translated texts. We selected three texts of the English Machine Translation Evaluation version (CREG-MT-eval) of the Corpus of Reading Comprehension Exercises (CREG), for which we produced three different translations: a manual translation and two automatic translations generated by two state-of-the-art neural machine translation engines, viz. DeepL and Google Translate. The experiment was conducted via a SurveyMonkey questionnaire, which 99 participants filled in. Participants were asked to read the translation very carefully after which they had to answer the comprehension questions without having access to the translated text. Apart from assessing comprehension, we posed additional questions to get information on the participants' perception of the machine translations. The results show that 74% of the participants can tell whether a translation was produced by a human or a machine. Human translations received the best overall clarity scores, but the reading comprehension tests provided much less unequivocal results. The errors that bother readers most relate to grammar, sentence length, level of idiomaticity and incoherence.

1 Introduction

Machine translation systems cannot guarantee that the text they produce will be fluent and coherent in both syntax and semantics. Erroneous words and syntax occur frequently in machine translated text, leaving the reader to guess parts of the intended message.

With the arrival of neural machine translation (NMT), however, the quality of machine translation has increased significantly (Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017; Van Brussel et al., 2018; Shterionov et al., 2018). As such, machine translation is becoming an attractive solution to deal with the increased need for translated content, which is reflected in new use cases such as business-to-consumer e-commerce and user-generated content (see Levin et al. (2017) for the Booking.com example). This could mean that, in the near future, readers will be more often confronted with 'raw' (unedited) MT output, which poses the question of how comprehensible such raw machine translated texts really are.

Different methods have been applied to test comprehensibility of machine translation output. Ageeva et al. (2015) used a gap-filling task to evaluate machine translated sentences for Basque-Spanish and Tatar-Russian, whereas Berka et al. (2011) used a quiz-based evaluation method to assess short machine translated passages using yes/no-questions for English-Czech.

Tomita et al. (1993) used the reading comprehension sections of TOEFL tests (Test of English as a Foreign Language) to assess different machine translation systems for English-Japanese. Jones et al. (2005) applied a proficiency test for Arabic on Arabic-English machine translations and showed that machine translations slowed down the respondents in answering the questions and that the accuracy was lower compared to human translations. Scarton and Specia (2016) studied the comprehensibility of several machine translation systems using reading comprehension tests, and used a set of human translations as a control. In their experiments participants did not achieve higher scores when reading the human translated texts. They also found a large variability across participants.

The work presented here is largely inspired by the work of Scarton and Specia (2016) in the sense that we started from their data set, although with a slightly adapted methodology.

Proceedings of the 40th Conference Translating and the Computer, pages 120–126,
London, UK, November 15-16, 2018.

We also compare the results of reading comprehension tests on both human translated and raw (unedited) machine translated texts, but we focus on two state-of-the-art neural machine translation engines, DeepL and Google Translate, whereas their work dates from the pre-NMT era. Our study differs from previous work in the sense that we also compare the results of the reading comprehension tests with information we gathered on the participants' perception of the translated texts.

2 Methodology

We adopted the methodology of Scarton and Specia (2016) and selected three texts from their English Machine Translation Evaluation version (CREG-MT-eval)[1] of the Corpus of Reading Comprehension Exercises (CREG). The selected English source texts were short, self-contained texts of approximately 200 words each. The three English texts were translated manually into Dutch by a master student of translation at Ghent University and translated automatically by means of two neural machine translation engines, viz. DeepL and Google Translate.

For each text we formulated five reading comprehension questions, which were either new questions or Dutch translations of the original questions taken from CREG-MT-eval. We limited ourselves to three question forms: wh-questions, literal questions and reorganisation questions. We decided to leave out yes/no questions and true/false questions for the simple reason that it is impossible to judge whether or not the participants just guessed the correct answer. Inference questions were also excluded from the questionnaire as it is hard to know whether a reader could not answer such questions because he/she did not comprehend the text or whether he/she did not possess the required world knowledge. An example of a wh-question used is 'What is needed to produce paper?', and a literal question would be, for instance, 'How much energy is saved when three sheets of recycled paper are used?' A reorganisation question obliges the reader to look for the answer in several parts of the text and 'Which city is the front-runner according to the atlas? Why?' is such a question.

The questionnaire was set up in SurveyMonkey and distributed via e-mail and Facebook. In total, 99 participants took part in the experiments. The participants were asked to read the translations very carefully after which they had to answer the reading comprehension questions without having access to the translated text. Each participant read two different texts, which could be either a human translation and a machine translation or two machine translations. In order to collect a comparable amount of data across all conditions, we randomized the order of the translations and the texts across participants. In total, each translated text was read by a minimum of 21 and maximum of 23 participants.

After the participants had answered the reading comprehension questions, we showed the text again to the participants and asked them to judge whether the text was a human or a machine translated text, to assign a global clarity score of 1–5 to each translated text, to indicate the passages they did not understand and to list the errors that bothered them while reading the text.

The same procedure was repeated for the second text. The questionnaire ended with some profile questions. There was no time limit imposed on the experiments. Most participants finished in approximately 15 minutes.

Ninety-five participants filled in the profile section of the questionnaire. The average age of the participants was 27 years old. Sixty-one of the participants were female, 33 male and 1 participant selected the category other. No less than 43 of the participants indicated that their current education or degree was related to the field of languages and only 8 participants were not currently pursuing higher education or had not obtained a degree. Four persons mentioned that they had never used a machine translation service. Of the remaining 91, 74 were positive in their

[1] Retrieved from https://github.com/carolscarton/CREG-MT-eval

attitude towards machine translation, but 52 made a comment, usually that the translation should always be checked afterwards or that they only use it in certain cases. Twenty participants indicated that they mostly use machine translation for the purpose of information gisting and 8 even mentioned that they prefer DeepL over Google Translate.

3 Results

We first report on the answers to the more general questions of the sections in which the participants had access to the texts when answering the questions and compare these results with the comprehension test in which the participants did not have the translated text at their disposal.

3.1 Man or machine?

For each text, the participants were asked to judge whether the text they had just read was a human or a machine translation and to justify their answers. The contingency table (Table 1) shows the actual labels of the conditions alongside the labels assigned by the participants. Of the 195 judgements, 144 were correct, which means that in 74% of the cases the participants were able to tell whether the Dutch translation was produced by a human or a machine. There were about the same number of wrong judgements in each of the conditions (18 for the Human Translations, 18 for Google Translate and 15 for DeepL).

Actual Condition	**Perceived condition**	
	Human	Machine
Human Translation	47	18
Google Translate	18	47
DeepL	15	50

Table 1: Contingency table with judgements per condition

The categories that were most often mentioned by the participants who correctly classified the human translated texts were ‘fluency’, ‘lack of mistakes’, ‘coherence’ and ‘idiomaticity’. The categories that were most often mentioned by the participants who correctly classified the machine translated texts were ‘grammatical mistakes’, ‘unidiomatic constructions’, ‘inconsistent translations’ and ‘repetitions’.

Grammatical mistakes that were referred to in the comments were subject-verb agreement problems (‘ik kookt’ instead of ‘ik kook’) and more complex structural problems. Examples of unidiomatic constructions that were given are ‘Dat is veel plezier voor mij’, which is a very literal translation of ‘That is a lot of fun for me’ or ‘Dat bevalt me erg leuk’ as translation for the phrase ‘I really like that’, which is in fact the result of blending two expressions ‘Dat bevalt me’ (En: ‘That pleases me’) and ‘Dat vind ik erg leuk’ (En: ‘I like that very much’). As examples of inconsistent translations participants mentioned the mix of the more formal ‘u’ and less formal ‘je’ as translations of the pronoun ‘you’ in Dutch and the inconsistent translation of terms such as ‘gerecycled papier’ and ‘gerecycleerd papier’ (for ‘recycled paper’), which were used interchangeably. An example of a repetition that was given is ‘gerecycled papier van oud papier’ (En: ‘recycled paper of old paper’), and a literal repetition in the DeepL translation ‘we gebruiken overal papier... we gebruiken papier’ as translation of ‘we use paper everywhere’.

3.2 Clarity scores

The participants were also asked to give an overall clarity score of 1-5 to the text they had read, with 1 being ‘completely incomprehensible’ and 5 being ‘perfectly comprehensible’. The

distribution of the clarity scores per text and per translation method is given in Figure 1.

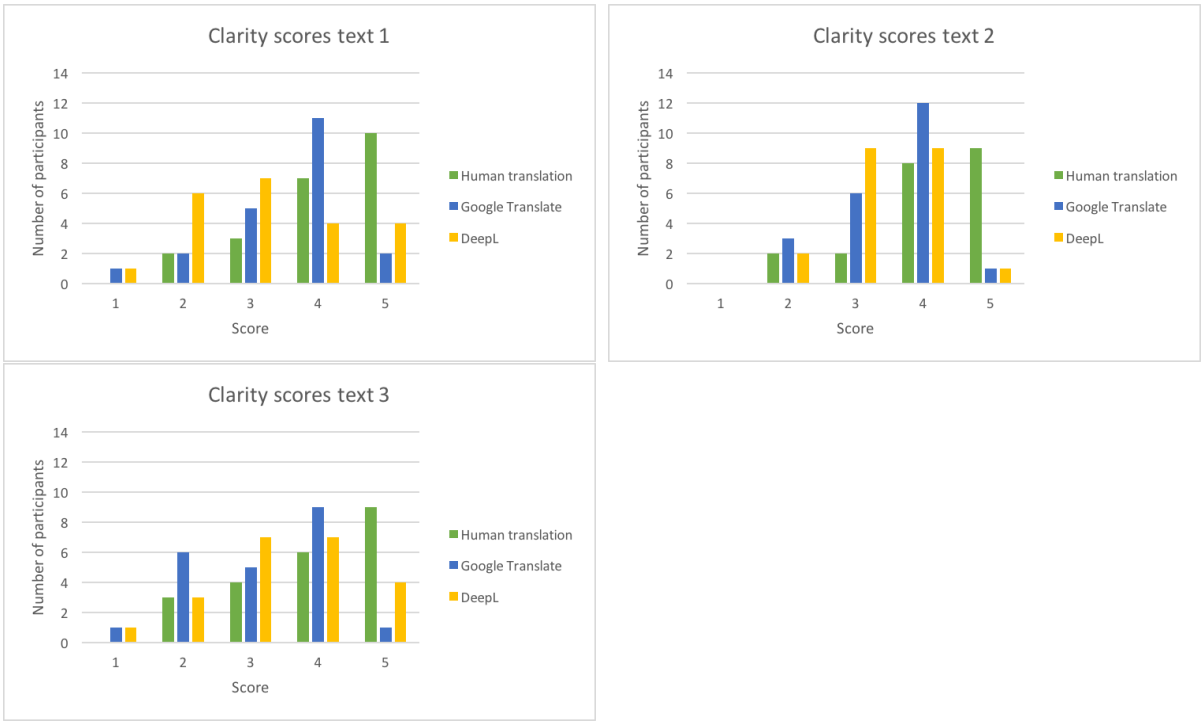

Figure 1: Distribution of the clarity scores per text and translation method

In general, the human translated texts were rated higher than the machine translated versions and they received more scores in the range of 4–5 and no single 1. As can be seen on the graphs presented in Figure 1, there is some variation across participants and the participants used the whole range of scores (1–5) for the machine translated versions of text 1 and text 3 and scores 2–5 for all human translations and the machine translated versions of text 2.

The averaged clarity scores are presented in Table 2. The human translated texts get an average score of 4.0–4.1, whereas the machine translations get average scores in the range of 3.1–3.5. Google Translate scores better for text 1 and DeepL better for text 3.

	Text 1	**Text 2**	**Text 3**
Human Translation	4.1	4.1	4.0
Google Translate	3.5	3.5	3.1
DeepL	3.2	3.4	3.5

Table 2: Averaged clarity scores per text and translation method

We also asked the participants to indicate the passages they had not understood. Some problematic passages were mentioned by multiple participants, but we see again large individual differences. Some passages indeed contained erroneous translations, others can be classified as unidiomatic expressions or repetitions. Some passages could, in retrospect, be linked to characteristics of the source text.

An example of an erroneous translation that hampered comprehension in the two machine translated texts was ‘My roommate’s family’, which was translated by Google Translate as ‘het huis van mijn kamergenoot’ (En: ‘the house of my roommate’), which changes the meaning of the source text fragment and was translated by DeepL as ‘mijn huisgenoot’ (En: ‘my house mate’), thus deleting the content word ‘family’ in the translation.

For text 3, a few participants did not cite any passages, but made more general comments that they had not understood the whole story (names, purpose, motive) and that the text lacked coherence.

3.3 Most irritating mistakes

When explicitly asked about the mistakes that bothered the participants when reading the texts, most remarks on the human translated texts related to stylistic issues. Some participants mentioned for text 1 that the sequence of short, simple sentences resulted in a staccato style of writing; some participants mentioned that the translations lacked cohesion because they did not contain enough discourse markers. Again, in retrospect, most of these remarks could be attributed to characteristics already present in the source text .

As for the machine translated texts, 12 out of 22 participants explicitly mentioned the wrong and non-sensical translation 'het huis van mijn kamergenoot woont in Berlijn' (En: 'The house of my roommate lives in Berlin'); other remarks related to other problems present in the NMT output such as repetitions, inconsistent translations, unidiomatic constructions, wrong anaphora, wrong tenses, wrong gender, the use of anglicisms. The remarks on the accumulation of short sentences and the lack of discourse markers were also raised for the machine translations.

3.4 Comprehension tests

To rate the comprehension test, we assigned scores of 1, 0.5 and 0 to each answer to the five comprehension questions, depending on the level of correctness of the answers (1 for completely correct answers, 0.5 for partially correct answers and 0 for incorrect answers). To receive a score of 0.5, the answer had to contain at least one element of the gold standard answer. The averaged comprehension scores per text and translation method are presented in Table 3.

To our surprise, the human translation only received the highest average comprehension score for text 1, and DeepL scored best for text 2 and text 3. It can also be noted that the results for text 2 are much lower than for the two other texts.

We examined the partial and erroneous answers in detail and only some of them could be attributed to erroneous translations. A possible explanation for the lower scores for text 2, is that most of the questions were about details and family members, which might have been much easier to answer if the text was displayed during the comprehension tests.

We came to the conclusion that the results of the comprehension tests can be (partly) explained by the experimental set-up. By not showing the texts during the comprehension tests, we test more recall than comprehensibility. It might well be that more effort is required to read a text that contains mistakes, and this increased effort might be the reason that participants remembered the content better.

	Text 1	**Text 2**	**Text 3**
Human Translation	3.4	2.4	3.1
Google Translate	3.0	1.6	3.3
DeepL	2.4	2.6	3.5

Table 3: Average comprehension score per text and translation method

4 Discussion

We found that 74% of all participants could correctly discern a human translation from a machine translation and 26% could not. The participants who could distinguish a human

translation from a neural machine translation usually relied on coherence, fluency, idiomaticity, clarity, sentence length and repetition to make this decision. It should be noted that the results might have been slightly influenced by the experimental set-up. The participants read either a human translated and a machine translated text or two machine translated texts. From some comments, however, we can assume that not all participants were aware of the fact that the translations they received could both be machine translations. Although we mentioned it in the introductory text in SurveyMonkey, for some reason participants expected two different conditions. In future work, we will either mention this more explicitly or – even better – not tell the participants at all that the texts they will read are translations.

A more plausible explanation is that the 26% of participants who could not distinguish machine translations from human translation are just not as sensitive to linguistic mistakes as the majority of the participants or even more tolerant towards textual disturbances caused by machine translation (Roturier, 2006). As a large part of the participants group was recruited amongst linguists, we checked whether linguistic background was a determining factor and this was not the case. We compared the percentage of correct judgements given by linguists and non-linguists and found no major differences. Also linguists and non-linguists assigned similar clarity scores.

The human translations obtained the best clarity scores, but when looking at the distribution of the scores, we observed some variation across participants. The mistakes that bother readers most have to do with problems in the machine translated output such as grammatical problems, repetitions, inconsistent translations and unidiomatic constructions, but also with stylistic issues such as short sentences, lack of coherence and missing discourse markers which were also present in the human translations. The latter issues could be attributed to characteristics that were already present in the source text.

As for the comprehension test, our results showed that the human translation was only rated best once, while DeepL proved itself to be the best for two of the three texts. This latter finding suggests that the machine translations are comprehensible for the language pair English-Dutch and the machine translation tools Google Translate and DeepL, although participants quoted certain passages that hindered comprehension. These results are in line with the findings of Scarton and Specia (2016), in which participants did not achieve higher scores on the comprehension tests for human translated documents either.

When comparing the overall clarity scores with the results of the reading comprehension tests we come to mixed conclusions. The human translations received the best overall clarity scores, but the reading comprehension tests provided less unequivocal results. This may be attributed to our decision not to display the text when taking the comprehension test. As suggested above, due to this decision, the focus of this study might have shifted more from comprehension towards recall. But it might well be that the clarity scores and the reading comprehension test assess different aspects of reading comprehension, which is known as a complex cognitive process.

5 Future work

Although reading comprehension tests and clarity scores provide useful insights in the amount of information that is retrieved and retained from a text, these methods do not tell anything about the underlying comprehension process. We assume that reading machine-translated text is a comprehension process that might be fundamentally different from the reading of normal, well-formed text as erroneous words and ungrammatical sentences occur (frequently) in machine-translated text, leaving the reader to guess parts of the intended message. This pilot study is part of a larger project, in which we will collect and analyse eye movements

of participants reading Dutch machine-translated text to investigate the impact of different categories of MT errors (syntactic versus semantic, function words versus content words, short-distance versus long-distance triggers of errors) on the underlying comprehension process.

Acknowledgments

This pilot study is part of the ArisToCAT project (Assessing The Comprehensibility of Automatic Translations)[2], which is a four-year research project (2017-2020) funded by the Research Foundation - Flanders (FWO) – grant number G.0064.17N.

References

Ekaterina Ageeva, Mikel L. Forcada, Francis M. Tyers, and Juan Antonio Prez-Ortiz. 2015. Evaluating machine translation for assimilation via a gap-filling task. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas, November. Association for Computational Linguistics.

Jan Berka, Martin Černý, and Ondřej Bojar. 2011. Quiz-Based Evaluation of Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 95(-1), January.

Douglas Jones, Edward Gibson, Wade Shen, Neil Granoien, Martha Herzog, Douglas Reynolds, and Clifford Weinstein. 2005. Measuring human readability of machine generated text: three case studies in speech recognition and machine translation. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 1009–1012. IEEE.

Pavel Levin, Nishikant Dhanuka, and Maxim Khalilov. 2017. Machine translation at Booking.com: Journey and lessons learned. In *Proceedings of the 20th Conference of the European Association for Machine Translation*, volume User Studies and Project/Product Descriptions, pages 81–86, Prague, Czech Republic. EAMT.

Johann Roturier. 2006. *An investigation into the impact of controlled English rules on comprehensibility, usefulness and acceptibility of machine-translated technical documentation for French and German users.* Ph.D. thesis, Dublin City University.

Carolina Scarton and Lucia Specia. 2016. A reading comprehension corpus for machine translation evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Dimitar Shterionov, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'Dowd, and Andy Way. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, May.

Masaru Tomita, Masako Shirai, Junya Tsutsumi, Miki Matsumura, and Yuki Yoshikawa. 1993. Evaluation of MT Systems by TOEFL. In *Proceedings of the Theoretical and Methodological Implications of Machine Translation (TMI-93)*.

Antonio Toral and Víctor 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*, pages 1063–1073, Valencia, Spain, April. Association for Computational Linguistics.

Laura Van Brussel, Arda Tezcan, and Lieve Macken. 2018. A fine-grained error analysis of NMT, SMT and RBMT output for English-to-Dutch. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3799–3804, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

[2]https://research.flw.ugent.be/en/projects/aristocat