Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 29–38
Marseille, 20-25 June 2022

GECO-MT: The Ghent Eye-tracking Corpus of Machine Translation

Toon Colman[1], Margot Fonteyne[1], Joke Daems[1], Nicolas Dirix[2], Lieve Macken[1]
[1]LT3, Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
[2]Department of Experimental Psychology, Ghent University
Henri Dunantlaan 2, 9000 Ghent, Belgium
{toon.colman, margot.fonteyne, joke.daems, nicolas.dirix, lieve.macken}@ugent.be

Abstract

In the present paper, we describe a large corpus of eye movement data, collected during natural reading of a human translation and a machine translation of a full novel. This data set, called GECO-MT (Ghent Eye-tracking Corpus of Machine Translation) expands upon an earlier corpus called GECO (Ghent Eye-tracking Corpus) by Cop et al. (2017). The eye movement data in GECO-MT will be used in future research to investigate the effect of machine translation on the reading process and the effects of various error types on reading. In this article, we describe in detail the materials and data collection procedure of GECO-MT. Extensive information on the language proficiency of our participants is given, as well as a comparison with the participants of the original GECO. We investigate the distribution of a selection of important eye movement variables and explore the possibilities for future analyses of the data. GECO-MT is freely available at https://www.lt3.ugent.be/resources/geco-mt.

Keywords: literary machine translation, quality assessment, reading behaviour, eye-tracking, corpus study

1. Introduction

The quality of machine translation (MT) output has increased greatly over the last decade, mainly thanks to a paradigm shift from *statistical* machine translation (SMT) systems to *neural* machine translation (NMT) systems (Wu et al., 2016). NMT generally outperforms SMT, as it is better able to account for (sentence) context and can map the meaning of words more finely. Both human evaluation methods and automatic metrics have shown NMT output to be more qualitative than SMT output (Bentivogli et al., 2016; Burchardt et al., 2017; Toral and Sánchez-Cartagena, 2017; Klubička et al., 2018; Van Brussel et al., 2018; Shterionov et al., 2018; Jia et al., 2019; Daems and Macken, 2019). Thanks to the increasing quality of MT, readers might be more often confronted with 'raw' MT output, without any post-editing[1] (Macken et al., 2020).

However, despite overall quality improvements, remarkable differences can be observed when comparing machine translations (MT) and human translations (HT), especially when considering more creative text types such as literary text. Webster et al. (2020) compared the Dutch HTs of four classic English novels with their MT versions, generated by Google Translate and DeepL (two NMT systems). They found that a large proportion of MT sentences contained errors. Using the SCATE (Smart Computer-aided Translation Environment) MT error taxonomy by Tezcan et al. (2017), they observed that the most frequent error types in their data set were (1) mistranslations, (2) coherence errors, and (3) style & register errors. These findings correspond closely to previous research by Tezcan et al. (2019) and Fonteyne et al. (2020) who both discussed the quality of Agatha Christie's novel *The Mysterious Affair at Styles*, translated by Google's NMT system from English into Dutch. Aside from errors, Webster et al. (2020) and Tezcan et al. (2019) also observed a lower level of lexical richness and cohesion in the MT novels compared to the HT versions. Although researchers have shown increasing interest in using NMT for literary translation (Toral and Way, 2018; Kuzman et al., 2019; Matusov, 2019), it is clear that many challenges remain. This, however, makes literary MT a suitable use case to study the effects of MT on text comprehension and reading behaviour, since MT errors are plenty and varied.

In the current project, we shift our focus from studying MT *output* (e.g., via MT error annotation) to studying the *reader*, the end user of the MT output. More specifically, we are interested in natural eye movement when people read MT text compared to HT text. To this end, we have collected a large corpus of eye-tracking data on both the Dutch HT and the Dutch MT of Agatha Christie's *The Mysterious Affair at Styles*. This corpus, called GECO-MT (Ghent Eye-tracking Corpus of Machine Translation), will allow us to investigate to what extent MT impacts the reading process. In future research, we will also be able to study which errors impact reading most, since a human annotator has marked and classified all errors in the MT version of the novel (Fonteyne et al., 2020). The GECO-MT data set builds upon the earlier GECO (Ghent Eye-Tracking Corpus) (Cop et al., 2017), which contains eye movement data of participants reading the same novel in the English original version and in the Dutch HT.

[1]The ArisToCAT project (Assessing the Comprehensibility of Automatic Translations) – which this study is part of – aims to evaluate the comprehensibility of 'raw' (unedited) MT output for readers who can only rely on the MT output.

2. Related work

Eye movement during reading consists of two basic components, namely (1) eye fixations and (2) eye saccades (Rayner, 1998; Rayner, 2009). *Fixations* are the instances in which the eyes remain relatively still and the reader extracts information from a piece of text. The average fixation duration is around 200 - 250 ms. *Saccades* are the actual movements of the eyes between subsequent fixations, during which the reader is functionally 'blind'. The average saccade amplitude is around 7 - 9 letter spaces and takes around 30 ms. The majority of more specific eye movement variables, such as the first fixation duration on a word, or the proportion of regressive (i.e., 'backwards') saccades during reading, are derived from these two basic components. Generally speaking, researchers find that "as text gets more difficult, fixations get longer, saccades get shorter, and more regressions are made" (Rayner, 2009). Several theoretical models have been put forward to more precisely explain the patterns found in eye movement data (Rayner and Reichle, 2010). The most influential is the E-Z Reader model (Reichle et al., 1998), which is a computational model that elegantly explains how lexical factors such as word frequency and word predictability predict fixation duration and word skipping.

Eye-tracking has previously been used to assess the quality of MT output. Doherty et al. (2010) first investigated whether MT quality is reflected in eye movement data. They found that when participants read MT sentences that were rated as poor by human evaluators, the number of fixations increased, as well as the average gaze duration (the sum of fixation durations on first-pass reading). The average fixation duration, however, was not affected. Stymne et al. (2012) used short MT texts instead of isolated sentences. They found no differences in eye movement between the HT and MT texts overall, but when zooming in on MT errors, differences emerged. In line with Doherty et al. (2010), the number of fixations and the gaze duration were higher for text fragments containing MT errors compared to correct MT output. Interestingly, gaze duration differed significantly between the different MT error categories which were based on the taxonomy of Vilar et al. (2006), with word order errors having the longest gaze duration, followed by incorrect or missing words. Kasperavičienė et al. (2020) used MT news articles and again found an increased number of fixations and gaze duration for MT errors compared to correct MT segments. The highest gaze durations and number of fixations were found for lexical errors, followed by linguistic morphological errors. These studies demonstrate that eye movement data can be useful to assess the readability of MT output and the severity of various types of MT errors.

In more applied research, eye-tracking has contributed to evaluating the usability of MT text compared to HT text or untranslated text. Doherty and O'Brien (2014) studied the usability of task instructions in technical support documentation for an online file storage system. Compared to the untranslated English instructions, Japanese MT was associated with longer task completion times and an increased number of eye fixations and fixation duration. Spanish, French, and German MT instructions were found to be equally usable to the original English text. Hu et al. (2020) compared the usability of translated subtitles for online educational videos. Raw MT subtitles were found to be equally usable as HT subtitles, but post-edited MT subtitles were estimated to be better (even outperforming the HT subtitles), based on lower average fixation durations. Finally, Guerberof Arenas et al. (2021) compared the usability of the Microsoft Word interface in different translation modalities. They found that when the interface was translated (into German, Japanese, or Spanish), average fixation durations during various tasks increased, compared to the untranslated English interface. There was no difference, however, between HT and MT.

It is clear that eye-tracking proves an interesting research method to study the effect of MT and MT errors on the reading process, as well as the usability of MT content in real life applications. When MT output becomes more difficult to read, it is expected to influence eye movement, resulting in longer fixation durations, shorter saccade amplitudes, a higher proportion of regressive saccades, and so forth (Rayner, 2009). Eye-tracking is an unobtrusive measure, meaning that no additional response or decision processes are mingled with the actual reading processes that are of interest (as opposed to self-report methods). Moreover, eye-tracking data has a high spatial and temporal resolution. Fixation locations can be determined with a spatial accuracy of 0.25 - 0.50 visual degrees and temporally, up to 2000 sample points per second can be collected (i.e., sampling rate of 2000 Hz). This makes eye-tracking a very fine-grained measurement technique. The advantages of eye-tracking methodology can be further leveraged by collecting very large amounts of data in big corpus studies (Kennedy and Pynte, 2005; Kliegl et al., 2006; Kuperman et al., 2010; Frank et al., 2013; Cop et al., 2017). Previous studies on eye movement and MT quality estimation (Doherty et al., 2010; Stymne et al., 2012; Kasperavičienė et al., 2020) used rather short text fragments, which decreases statistical power and the ability to detect small effects. By collecting eye movements from the reading of a *full novel*, an extensive corpus can be created that lends itself to comprehensive statistical analyses. Moreover, variables of interest such as MT errors but also word frequency and word length vary naturally within the large corpus instead of being manipulated in a contrived experimental design.

In the present study, we expand the existing GECO (Cop et al., 2017), thereby creating the first eye movement corpus of MT reading of an entire novel, called

GECO-MT. In the following paragraphs, we will first describe the materials used for GECO-MT, as well as the data collection procedure. Then we will take a closer look at a selection of eye movement variables that are informative of MT reading, by presenting descriptive statistics and visualizations. In future work, these variables will be of vital importance to study the effects of MT and MT errors on the reading process.

3. Method

3.1. Textual materials

In the present study, we used a Dutch HT and a Dutch MT of the originally English detective novel *The Mysterious Affair at Styles* by Agatha Christie. The Dutch HT is identical to the Dutch materials used in the original GECO project (Cop et al., 2017). The authors originally selected this text because of the above-average reading ease, and close similarity between the word frequency distribution of the novel and the word frequency distribution in natural language use. Selecting a more difficult (e.g., more poetic) literary text could lead to difficulties in comparing the HT and MT due to a high variability in translation options (i.e., multiple possible translations). The Dutch MT is taken from the MT error analysis study by Fonteyne et al. (2020). The authors generated an English-into-Dutch machine translation, using the NMT system Google Translate, in May 2019[2]. The Dutch HT contains 5,190 sentences, and 59,716 words. The Dutch MT contains 5,276 sentences and 58,039 words. Both the HT and the MT were divided into 672 aligned paragraphs, to be presented one at a time on a computer screen during data collection.

Fonteyne et al. (2020) enriched the Dutch MT materials, by providing fine-grained MT error annotations using the SCATE MT error taxonomy (Tezcan et al., 2017). This taxonomy aids in categorizing MT errors, based on the well known distinction between accuracy and fluency errors (Figure 1). Using further subcategories, the MT errors are classified in a maximum of three hierarchical levels (e.g., accuracy → mistranslation → semantically unrelated). 66.1% of sentences in the Dutch MT contain at least one MT error. 58.8% of total MT errors are fluency errors, and 41.2% are accuracy errors. When observing level 2 of the MT error hierarchy, the most represented error categories are mistranslation (34.7%), coherence (30.5%), and style & register (15.8%) errors. An example of a *mistranslation* error in the MT materials is "Come and be refreshed!" translated into "Kom en word vernieuwd!" (literal back translation: "Come and be renewed!"). An example of a *coherence* error is "We had a good yearn about old times." translated into "We hadden een goed garen over oude tijden." ("een goed garen" means something like "a good ball of wool", making the sentence illogical and confusing). An example of a *style & register* error is "Where is tea today?" translated into "Waar is thee vandaag?" (literal translation of an expression that is idiomatic in English but not in Dutch). For full information on the annotation procedure and results, please refer to Fonteyne et al. (2020).

In principle, only fluency errors can influence eye movement, since participants only have access to the translation and not the original source text. Accuracy errors imply a comparison between the source text and the translation. However, Fonteyne et al. (2020) found that accuracy problems often lead to fluency problems (e.g., mistranslations leading to logical problems). Therefore, we might still be able to investigate effects that accuracy errors have on reading behaviour.

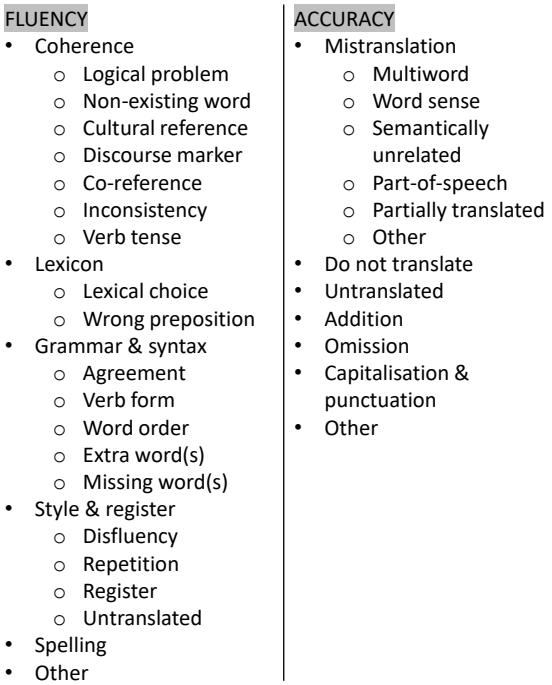

Figure 1: Visualization of the SCATE MT error taxonomy, taken from Tezcan et al. (2019). All MT errors present in the MT text materials were categorized in a maximum of three hierarchical levels. Observing level 2 of the hierarchy, the most frequent error types are (1) mistranslations, (2) coherence errors, and (3) style & register errors.

3.2. Participants

We recruited 20 Dutch-speaking participants (18 female, M_{age} = 21.40, SD_{age} = 2.21), all of which were enrolled in a bachelor's or master's program in applied language studies. All participants had normal or corrected-to-normal eyesight, and none of the partici-

[2]In 2022, state-of-the-art NMT systems are based on transformers, as opposed to recurrent neural networks in 2019. This is not problematic for the research scope of GECO-MT, as the corpus will be used to assess the effect of various MT error types on reading behaviour. To this end, we simply need MT output that contains a sufficient number and variety of MT errors.

Instrument	GECO	GECO-MT	t-value [df]	p-value
Dutch LexTALE (%)	92.43 [6.33]	90.75 [6.63]	-0.81 [37.00]	.456
Dutch classical LDT (%)	80.19 [5.40]	84.20 [7.34]	1.93 [34.71]	.186
Dutch spelling (%)	83.15 [7.81]	91.50 [6.53]	3.61 [35.15]	.006 **
English LexTALE (%)	75.63 [12.86]	80.44 [11.54]	1.23 [36.06]	.456
English classical LDT (%)	56.83 [11.11]	65.57 [12.04]	2.32 [35.97]	.104
English spelling (%)	69.92 [8.73]	77.50 [7.70]	2.87 [35.87]	.035 *

Table 1: *M* [and *SD*] of the percentage scores for all language proficiency tests, both for the original GECO (Cop et al., 2017) and for GECO-MT. Unpaired, two-sided t-test comparisons between both corpora are also provided. P-values were corrected for multiple testing, using the Holm (1979) method.
$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

pants had any diagnosed reading or language impairments. We also ensured that no participant had previously read Agatha Christie's *The Mysterious Affair at Styles*. At the start of the study, all participants signed an informed consent form and at the end of the study, each was rewarded 125 euros.

We collected extensive Dutch and English language proficiency data on all our participants (also see 3.4 Procedure). The language proficiency instruments were very similar to those in the original GECO-project (Cop et al., 2017). We only omitted the LEAP-Q self-report questionnaire for subjective language exposure (Marian et al., 2007), since it was of limited relevance to us. Both for Dutch and for English we administered an unspeeded and a speeded lexical decision task (LDT). In a typical LDT, participants are presented letter strings on a computer screen and have to respond (using response buttons) whether they think it is an existing word or a non-word. LDTs are widely used to measure vocabulary size and general language proficiency. First, we used the Dutch and English LexTALE (Lemhöfer and Broersma, 2012), which is a standardized unspeeded LDT and second, we used a Dutch and English speeded LDT with the same word and non-word stimuli as Cop et al. (2017). We also assessed spelling proficiency, using the Gl&schr spelling test (De Pessemier and Andries, 2009) for Dutch, and the 'green' spelling list from the WRAT 4 (Wilkinson and Robertson, 2006) for English.

The results of all our language proficiency tests are summarized in table 1. Means and standard deviations of the percentage scores per instrument are given, as well as a comparison with the results from the original GECO-project (Cop et al., 2017). The participants in GECO-MT scored significantly higher on the Dutch and English spelling tests, compared to the participants in the original GECO.

3.3. Apparatus

During data collection, the text materials were presented per paragraph on a computer screen of 40.5 x 31.0 cm. Participants were seated approximately 95.0 cm from the screen. The text was set on a light gray background, in black 14-point Courier New font with triple interline spacing. An Eyelink Portable Duo eye-tracking system (SR Research) was placed 45.0 cm from the right eye of the participant. The eye-tracker was used to measure eye movement of the right eye only, at a sampling rate of 2000 Hz. The presentation of the text materials and recording of the eye-tracking data were programmed using the Experiment Builder software package (SR Research, 2020). The LDTs to measure language proficiency were programmed using the PsychoPy2 software package (Peirce et al., 2019). To mask any distracting background noises during data collection, participants wore noise-cancelling wireless headphones playing Brownian noise audio.

3.4. Procedure

To collect our data, we conducted a reading experiment, closely mirroring the experiment in the original GECO-project (Cop et al., 2017). The full procedure received ethical approval by the institutional review board at Ghent University. Each participant attended four experimental sessions of approximately two and a half hours apiece. The four sessions were spread over a maximum period of three weeks, leaving minimum one day between every session. Participants read one quarter of Agatha Christie's *The Mysterious Affair at Styles* in each session[3], alternating between the HT and the MT. Half of the participants (those with an uneven participant number) started with the HT in session one, while the other half started with the MT. Thus, the conditions were counterbalanced so to avoid any order effects in the reading data. Participants were not informed of the experimental manipulation beforehand. They were simply told at the start of the experiment that we would investigate natural eye movement during the reading of a translated novel, and the effects of translation quality on reading.

Before reading commenced, the eye-tracking system was calibrated using a 9-point calibration procedure. Participants were then presented the text materials (HT or MT, depending on the experimental condition) one paragraph at a time. Using the computer keyboard, par-

[3]Chapters 1-4 in session one, 5-7 in session two, 8-10 in session three, and 11-13 in session four.

ticipants could progress to the next paragraph. After each paragraph, a drift-check was carried out by the eye-tracking system. If the experimenter judged that drift was too high (exceeding 0.5 visual degrees), the calibration procedure was repeated. Every 15 minutes, participants were allowed to take a break, after which the eye-tracking system was also re-calibrated. A chin rest was used to aid the participant in minimizing head movement.

At the end of each chapter, multiple-choice comprehension questions concerning the plot of the chapter were filled out, using pen and paper. The primary motivation for these questions was to give the participants an incentive to read the novel attentively. We also tested the *subjective* comprehension, at the end of each experimental session. Participants were instructed to read summaries of all chapters in the session, and judge how well the contents of the summaries aligned with the contents of the reading materials, by assigning a percentage score. As the experiment was conducted over four separate days, at the start of each session (except session one), participants again read the summaries of the *preceding* session to refresh their memory of the story.

Finally, as discussed in section 3.2, we assessed the Dutch and English language proficiency of each participant. At the end of session one, the Dutch and English LexTALE (Lemhöfer and Broersma, 2012) were administered. Concluding session two, the Dutch Gl&schr spelling test (De Pessemier and Andries, 2009) and the English WRAT 4 spelling test (Wilkinson and Robertson, 2006) were completed. At the end of sessions three and four, the Dutch and English classical LDTs were administered, respectively. The results of these proficiency tests are presented in table 1.

4. Results

4.1. Text comprehension

Before proceeding to the eye movement data, we will shortly discuss the text comprehension results. Firstly, the percentage scores on the multiple-choice comprehension questions were similar after reading HT (M = 85.67%, SD = 11.26%) versus MT (M = 82.03%, SD = 15.73%), $t(19) = 1.25$, $p = .113$. Thus, the 'objective' text comprehension did not differ between the HT and MT conditions. We did, however, find a significant difference in the subjective comprehension scores. After reading the HT, participants judged the similarity of the text materials to the chapter summaries to be higher (M = 86.50%, SD = 10.92%) than after reading MT (M = 81.07%, SD = 7.48%), $t(19) = 3.18$, $p = .002$. The subjective text comprehension was therefore higher in the HT condition, compared to the MT condition.

4.2. Pre-processing of eye movement data

All eye movement data were pre-processed, using the Data Viewer software package (SR Research, 2019). Firstly, eye fixations with a duration below 100 ms were removed since they are not thought to reflect any cognitive processing (Sereno and Rayner, 2003). Then, we exported a large number of eye movement variables on both (a) the paragraph level and (b) the word level. Using these data, comparisons between the HT and MT conditions can be made, at the level of paragraphs and at the level of words (and sequences of words). GECO-MT contains 20 eye movement variables at the paragraph level, and 50 variables at the word level. The latter contains all variables that are present in GECO (Cop et al., 2017). Paragraph level data were not included in the original GECO. The GECO-MT data are freely available online, along with documentation explaining each variable in the corpus.

After exporting the variables, further data processing and analyses were performed using R (R Core Team, 2020). We removed outlier observations, per participant and per variable. For each eye movement variable, we calculated the participant mean and standard deviation. Observations deviating more than 2.5 standard deviations from the mean were omitted. For all data visualizations, we also $\log_{10}$-transformed our data, which helps to better approach a normal distribution. It might be advisable to run a log-transformation when performing statistical tests on GECO-MT data, since the validity of some statistical techniques depends on the normality assumption. Nevertheless, all descriptive statistics (means and standard deviations) are presented on the non-transformed data, since this allows for easy interpretation in the original measurement units (e.g., milliseconds, visual degrees, etc.).

4.3. Distribution of selected eye movement variables

4.3.1. Paragraph level

To describe our paragraph level data, we have made a selection of four variables, based on the measures that were previously investigated in studies on MT quality estimation (Doherty et al., 2010; Stymne et al., 2012; Kasperavičienė et al., 2020). These variables are (1) *reading duration* (RD) of the current paragraph, (2) *number of fixations* (NF) on the current paragraph, (3) *average fixation duration* (AFD) of all fixations on the current paragraph, and (4) *average saccade amplitude* (ASA) of all saccades on the current paragraph.

Variable	Dutch HT	Dutch MT
RD (s)	20.77 [5.60]	21.70 [6.18]
NF (count)	77.61 [19.62]	80.11 [21.31]
AFD (ms)	221.51 [19.97]	225.10 [19.67]
ASA (deg.°)	3.36 [0.55]	3.25 [0.53]

Table 2: M [and SD] of the paragraph level variables *reading duration* (RD), *number of fixations* (NF), *average fixation duration* (AFD), and *average saccade amplitude* (ASA), after outlier removal.

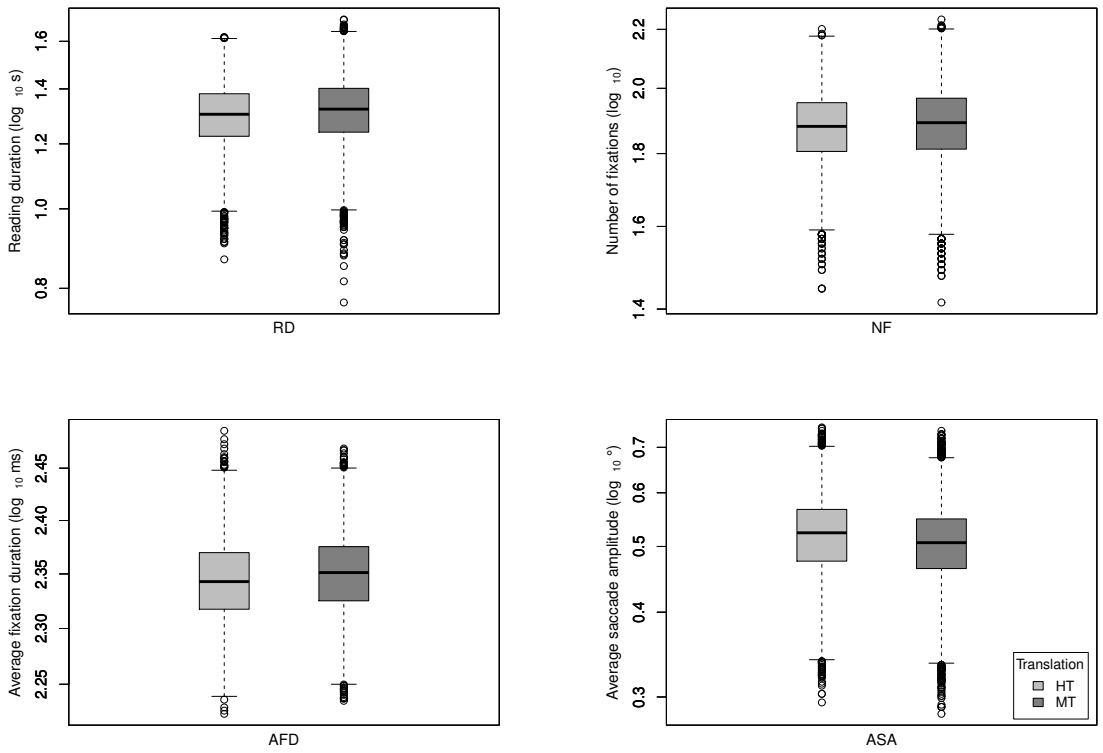

Figure 2: Boxplots visualizing the distribution of the paragraph level variables *reading duration* (RD), *number of fixations* (NF), *average fixation duration* (AFD), and *average saccade amplitude* (ASA), after outlier removal and $\log_{10}$-transformation.

Means and standard deviations of all four variables are presented in table 2. In line with earlier research, the average paragraph RD is slightly higher for MT reading compared to HT reading, as well as the NF, and the AFD. The ASA is slightly shorter for MT reading compared to HT reading, which is also as expected, if the MT is more difficult to read than the HT. Note, however, that the *overall* distribution of these variables will not necessarily reveal big differences between HT and MT reading. Possibly, differences emerge more clearly when zooming in on MT segments that contain errors, as previously found by Stymne et al. (2012).

In figure 2, we visualize the distribution of the four variables in the HT and the MT after $\log_{10}$-transformation. Although the transformation notably helps to normalize the distribution, the boxplots demonstrate that the distributions are still not entirely normal. For example, the distributions of RD and NF are slightly skewed to the left (i.e., more low values) while the distributions of AFD and ASA are more symmetrical. This characteristic of the reading data should be taken into account when performing analyses on GECO-MT.

4.3.2. Word level

At the word level, we have chosen to describe the same five variables that were discussed in Cop et al. (2017), in order to allow for direct comparison. The selected variables are (1) *first fixation duration* (FFD), which is the duration of the first fixation on the current word, (2) *single fixation duration* (SFD), the fixation duration for the subset of words that were fixated only once, (3) *gaze duration* (GD), which is the sum of all fixation durations in the first-pass reading, before the eyes move out of the word, (4) *total reading time* (TRT), the sum of all fixation durations on the current word, including re-fixations after regressions, and finally (5) *go-past time* (GPT), which is the summed fixation duration from when the current word is first fixated until the eyes move to the right of the current word, thus including regressions to previous words.

In table 3, we present the means and standard deviations of all five variables, along with the Dutch HT from the original GECO[4] (Cop et al., 2017). Firstly, the average reading times in GECO-MT seem to be slightly higher than in GECO. This may be due to the different

[4]The reported statistics do not correspond exactly to those reported in Cop et al. (2017) due to slightly different preprocessing choices.

Variable	Dutch HT (GECO)	Dutch HT (GECO-MT)	Dutch MT
FFD (ms)	209.05 [64.88]	215.27 [65.62]	218.49 [67.56]
SFD (ms)	210.18 [64.31]	216.12 [65.09]	219.52 [67.16]
GD (ms)	217.64 [71.14]	224.60 [73.04]	228.86 [75.75]
TRT (ms)	225.86 [75.79]	233.50 [79.09]	239.80 [82.81]
GPT (ms)	268.79 [132.13]	275.16 [144.00]	293.67 [172.16]

Table 3: *M* [and *SD*] of the word level variables *first fixation duration* (FFD), *single fixation duration* (SFD), *gaze duration* (GD), *total reading time* (TRT), and *go-past time* (GPT), after outlier removal. We have included the means from the original GECO (Cop et al., 2017) to allow for comparison with GECO-MT.

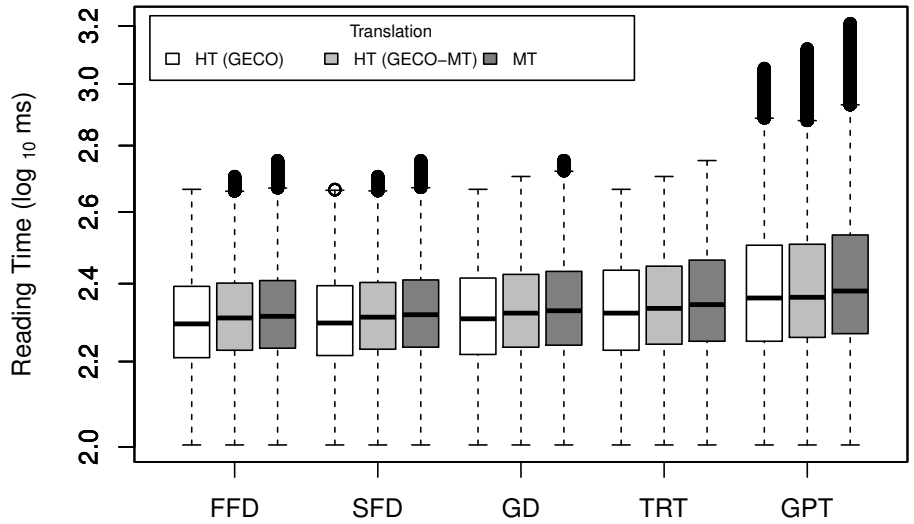

Figure 3: Boxplots visualizing the distribution of the word level variables *first fixation duration* (FFD), *single fixation duration* (SFD), *gaze duration* (GD), *total reading time* (TRT), and *go-past time* (GPT), after outlier removal and $\log_{10}$-transformation. We have included boxplots from the original GECO (Cop et al., 2017) to allow for comparison with GECO-MT.

participant groups that were used. Possibly, the participants in GECO-MT had a slightly 'slower' reading style than the participants in GECO. Secondly, reading times seem to be slightly higher for the MT compared to the HT, especially the *later* reading measures (e.g., measures including re-fixations after regressive saccades), such as the TRT and GPT. As with the paragraph level data, this suggests lower readability of the MT compared to the HT. Note again, however, that if there are any real differences between the eye movement data in the MT compared to the HT, they are expected to emerge most clearly when zooming in on the MT segments that contain errors.

The boxplots in figure 3 compare the distribution of all five variables described above, between the Dutch HT in GECO (Cop et al., 2017), the HT in GECO-MT, and the MT in GECO-MT. Again, despite log-transformation, the distribution of the reading times is not entirely normal. The variables are skewed to the right (i.e., more high values) which is typical of reading data (Frank et al., 2013; Cop et al., 2017).

5. Discussion

In this article we introduced GECO-MT, which is a large corpus of eye movement data, collected during the reading of a HT and MT of a full novel (*The Mysterious Affair at Styles* by Agatha Christie). GECO-MT will be a valuable resource to investigate the effects of unedited MT on reading behaviour (i.e., eye movement). Since the MT of the novel was enriched with fine-grained MT error annotations (Fonteyne et al., 2020), it will also be possible to determine the effect of specific types of MT errors.

Previous research already provided insight in the effects of MT and MT errors on eye movement (Doherty et al., 2010; Stymne et al., 2012; Kasperavičienė et al., 2020; Doherty and O'Brien, 2014; Hu et al., 2020; Guerberof Arenas et al., 2021). However, these studies used rather short text fragments. GECO-MT is collected on the reading of a *full novel* (HT and MT) and therefore contains a very large number of eye movement observations. This will significantly increase the statistical power of investigations in the effect of MT

on reading, and the ability to detect subtle effects of MT errors. The new corpus forms an extension of the earlier GECO (Cop et al., 2017) which contains eye movement data of participants reading the same novel in the English original version, and in the Dutch HT. As the Dutch HT is identical in GECO and GECO-MT, these subsets of both corpora can be considered equivalent. We found that the word level reading times follow a similar distribution in GECO and GECO-MT (see figure 3), suggesting that the data sets are indeed comparable. We do see however, that the reading times for the Dutch HT in GECO-MT are slightly higher than in GECO, suggesting that the participant group in GECO-MT had a slightly 'slower' overall reading style.

It is important to note that both corpora used independent samples of participants with a different educational background. The participants in GECO were students in psychology, while the participants in GECO-MT were students in applied language studies. In both corpora, the language proficiency of the participants was extensively tested, both for Dutch and English (see table 1). The results of these assessments reveal that the participants in GECO-MT overall have a slightly higher language proficiency than the participants in GECO. Therefore, it might be difficult to directly compare reading data from GECO (e.g., the English source text) and GECO-MT (e.g., the Dutch MT). Researchers should at least statistically control for the differences in language proficiency, as an effort to rule out confound effects.

GECO-MT contains data at two levels of analysis, (a) the paragraph level and (b) the word level. The original GECO (Cop et al., 2017) only contains the latter, which indeed provides the most fine-grained insight in the eye movement data. We have chosen to still include the paragraph level, as it contains some more general 'overview' measures that can be informative of MT reading. It will be interesting for example, to test the effect of MT on paragraph reading duration, or the total number of fixations on a paragraph. Moreover, the paragraph level measures are very robust to slight deviations in the calibration of the eye-tracking system, since the computation of most paragraph level variables does not depend on high spatial accuracy. The word level provides a more detailed look at the data, making it possible to directly compare subsets of words with certain characteristics, such as mistranslation errors, versus spelling errors. Thus, the word level lends itself best to compare the effect of the specific MT error types in the SCATE taxonomy (Tezcan et al., 2017).

Whether studying the paragraph level or the word level, it is expected that any differences in eye movement will emerge most clearly when zooming in on MT errors, as previously found by Stymne et al. (2012). At the paragraph level, we might compare paragraphs with more MT errors and paragraphs with less errors, while at the word level, we might compare sequences of words with errors and sequences without errors.

We explicitly chose to not provide any statistical hypothesis tests of the reading data in this article. We could have provided a simple t-test comparison between the participant means in the HT versus the MT – similarly to the language proficiency and text comprehension data – but this would do no justice to the complexity of the reading data, and might even lead to misleading conclusions. A t-test requires so-called *independence of observations*, meaning that the reading time variables have to be aggregated to *one* average per participant. This results in too much loss of information, since the reading times are influenced by a multitude of variables such as MT errors, but also part-of-speech, word length, and more. Besides, we also want to control for differences in participant language proficiency, as well as context effects, such as how far into the novel the reader has progressed.

In future studies, we will analyse the data from GECO-MT using the linear mixed effects (LME) model. The LME approach allows the reading data to be analysed in its non-aggregated form, while statistically controlling for word variables, as well as participant and context variables. We expect that MT errors will lead to an increased number of fixations, longer reading durations, and shorter saccade amplitudes. Moreover, we expect that some MT error types (e.g., mistranslations) will be more strongly implicated than other error types (e.g., spelling mistakes). The E-Z reader model (Reichle et al., 1998) can help with theoretically explaining which cognitive sub-processes are specifically affected by the MT errors. We also think that language proficiency, and especially English proficiency, could mitigate the effects of MT on reading behaviour. Even though the participants did not have access to the English source text, it is possible that some types of MT errors are more 'transparent' for participants with a high command of English. Finally, we expect that any effects of MT on reading behaviour might be especially pronounced at the beginning of a reading session. As the participant makes progress in the novel, knowledge of the plot and characters might facilitate reading of the MT text.

6. Acknowledgements

This study is part of the ArisToCAT project (Assessing The Comprehensibility of Automatic Translations), which is a research project funded by the Research Foundation – Flanders (FWO) – grant number G.0064.17N.

7. Bibliographical References

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (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.

Burchardt, A., Macketanz, V., Dehdari, J., Heigold, G., Peter, J.-T., and Williams, P. (2017). A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics*, 108:159–170, June.

Cop, U., Dirix, N., Drieghe, D., and Duyck, W. (2017). Presenting GECO : an eyetracking corpus of monolingual and bilingual sentence reading. *BEHAVIOR RESEARCH METHODS*, 49(2):602–615.

Daems, J. and Macken, L. (2019). Interactive adaptive SMT versus interactive adaptive NMT: a user experience evaluation. *Machine Translation*, 33(1):117–134, Jun.

De Pessemier, P. and Andries, C. (2009). *Gl&Schr: Test voor Gevorderd Lezen en Schrijven*. Garant, Leuven/Apeldoorn.

Doherty, S. and O'Brien, S. (2014). Assessing the usability of raw machine translated output: A user-centered study using eye tracking. *International Journal of Human–Computer Interaction*, 30(1):40–51.

Doherty, S., O'Brien, S., and Carl, M. (2010). Eye tracking as an MT evaluation technique. *Machine Translation*, 24(1):1–13.

Fonteyne, M., Tezcan, A., and Macken, L. (2020). Literary machine translation under the magnifying glass : assessing the quality of an NMT-translated detective novel on document level. In Calzolari, Nicoletta and Béchet, Frédéric and Blache, Philippe and Choukri, Khalid and Cieri, Christopher and Declerck, Thierry and Goggi, Sara and Isahara, Hitoshi and Maegaard, Bente and Mariani, Joseph and Mazo, Hélène and Moreno, Asuncion and Odijk, Jan and Piperidis, Stelios, editor, *12th International Conference on Language Resources and Evaluation Conference (LREC 2020), Proceedings*, pages 3783–3791. European Language Resources Association (ELRA).

Frank, S. L., Fernandez Monsalve, I., Thompson, R. L., and Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4):1182–1190, December.

Guerberof Arenas, A., Moorkens, J., and O'Brien, S. (2021). The impact of translation modality on user experience: an eye-tracking study of the Microsoft Word user interface. *Machine Translation*, 35(2):205–237, June.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Hu, K., O'Brien, S., and Kenny, D. (2020). A reception study of machine translated subtitles for MOOCs. *Perspectives*, 28(4):521–538.

Jia, Y., Carl, M., and Wang, X. (2019). Post-editing neural machine translation versus phrase-based machine translation for English–Chinese. *Machine Translation*, 33(1):9–29, Jun.

Kasperavičienė, R., Motiejūnienė, J., and Patašienė, I. (2020). Quality assessment of machine translation output: cognitive evaluation approach in an eye tracking experiment. *Texto Livre*, 13(2):271–285, Jul.

Kennedy, A. and Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, 45(2):153–168.

Kliegl, R., Nuthmann, A., and Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, 135(1):12–35.

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

Kuperman, V., Dambacher, M., Nuthmann, A., and Kliegl, R. (2010). The effect of word position on eye-movements in sentence and paragraph reading. *Quarterly Journal of Experimental Psychology*, 63(9):1838–1857.

Kuzman, T., Špela Vintar, and Arčan, M. (2019). Neural machine translation of literary texts from English to Slovene. In *Proceedings of the Qualities of Literary Machine Translation*, pages 1–9, Dublin, Ireland, 19 August. European Association for Machine Translation.

Lemhöfer, K. and Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, 44(2):325–343, June.

Macken, L., Fonteyne, M., Tezcan, A., and Daems, J. (2020). Assessing the Comprehensibility of Automatic Translations (ArisToCAT). In Mikel L. Forcada, André Martins, editor, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT2020)*, pages 485–486. European Association for Machine Translation (EAMT).

Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, 50(4):940–967.

Matusov, E. (2019). The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland, 19 August. European Association for Machine Translation.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., and Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203, February.

R Core Team, (2020). *R: A Language and Environment*

for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rayner, K. and Reichle, E. D. (2010). Models of the reading process. *WIREs Cognitive Science*, 1(6):787–799.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62(8):1457–1506.

Reichle, E. D., Pollatsek, A., Fisher, D. L., and Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, 105(1):125–157.

Sereno, S. C. and Rayner, K. (2003). Measuring word recognition in reading: eye movements and event-related potentials. *Trends in Cognitive Sciences*, 7(11):489–493.

Shterionov, D., Superbo, R., Nagle, P., Casanellas, L., O'Dowd, T., and Way, A. (2018). Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, 32(3):217–235, May.

SR Research, (2019). *EyeLink Data Viewer 4.1.1 [Computer software]*. SR Research Ltd., Mississauga, Ontario, Canada.

SR Research, (2020). *SR Research Experiment Builder 2.3.1 [Computer software]*. SR Research Ltd., Mississauga, Ontario, Canada.

Stymne, S., Danielsson, H., Bremin, S., Hu, H., Karlsson, J., Lillkull, A. P., and Wester, M. (2012). Eye tracking as a tool for machine translation error analysis. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1121–1126, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Tezcan, A., Hoste, V., and Macken, L. (2017). SCATE taxonomy and corpus of machine translation errors. In Gloria Corpas Pastor et al., editors, *Trends in E-tools and resources for translators and interpreters*, volume 45 of *Approaches to Translation Studies*, pages 219–244. Brill — Rodopi.

Tezcan, A., Daems, J., and Macken, L. (2019). When a 'sport' is a person and other issues for NMT of novels. In Hadley, James and Popović, Maja and Afli, Haithem and Way, Andy, editor, *Proceedings of the Qualities of Literary Machine Translation*, pages 40–49. European Association for Machine Translation.

Toral, A. and Sánchez-Cartagena, V. M. (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.

Toral, A. and Way, A. (2018). What level of quality can neural machine translation attain on literary text? In Joss Moorkens, et al., editors, *Translation Quality Assessment*, Machine Translation: Technologies and Applications, pages 263–287. Springer International Publishing AG.

Van Brussel, L., Tezcan, A., and Macken, L. (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).

Vilar, D., Xu, J., D'haro, L., and Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy. European Language Resources Association (ELRA).

Webster, R., Fonteyne, M., Tezcan, A., Macken, L., and Daems, J. (2020). Gutenberg goes neural : comparing features of Dutch human translations with raw neural machine translation outputs in a corpus of English literary classics. *INFORMATICS-BASEL*, 7(3):21.

Wilkinson, G. S. and Robertson, G. J., (2006). *Wide Range Achievement Test 4 professional manual*. Psychological Assessment Resources, Lutz, Florida.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation.