http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 13, n. 2, mai.-ago. 2020 – ISSN 1983-3652
DOI:

QUALITY ASSESSMENT OF MACHINE TRANSLATION OUTPUT: COGNITIVE EVALUATION APPROACH IN AN EYE TRACKING EXPERIMENT

AVALIAÇÃO DA QUALIDADE DA PRODUÇÃO DE TRADUÇÃO AUTOMÁTICA: ABORDAGEM DE AVALIAÇÃO COGNITIVA EM UM EXPERIMENTO COM RASTREAMENTO OCULAR

Ramunė Kasperavičienė
Kaunas University of Technology, Lituânia
ramune.kasperaviciene@ktu.lt

Jurgita Motiejūnienė
Kaunas University of Technology, Lituânia
jurgita.motiejuniene@ktu.lt

Irena Patašienė
Kaunas University of Technology, Lituânia
irena.patasiene@ktu.lt

ABSTRACT: Despite fast development of machine translation, the output quality is less than acceptable in certain language pairs. The aim of this paper is to determine the types of errors in machine translation output that cause comprehension problems to potential readers. The study is based on a reading task experiment using eye tracking and a retrospective survey as a complementary method to add more value to the research as eye tracking as a method is considered to be problematic and challenging (O'BRIEN, 2009; ALVES et al., 2009). The cognitive evaluation approach is used in an eye tracking experiment to determine the complexity of the errors in the English–Lithuanian language pair from easiest to hardest as seen by the readers of a machine-translated text. The tested parameters – gaze time and fixation count – demonstrate that a different amount of cognitive effort is required to process different types of errors in machine-translated texts. The current work aims at contributing to other research in the Translation Studies field by providing the analysis of error assessment of machine translation output.
KEYWORDS: machine translation; cognitive evaluation approach; translation error(s); eye tracking; acceptability.

RESUMO: Apesar do rápido desenvolvimento da tradução automática, a qualidade do texto produzido é bastante pobre em algumas combinações linguísticas. O objetivo deste artigo é determinar os tipos de erros na produção de tradução automática que acarretam dificuldades de compreensão para os potenciais leitores. O estudo é baseado em um experimento que utiliza rastreamento ocular e um questionário retrospetivo como método complementar de forma a acrescentar mais valor à pesquisa, visto que o rastreamento ocular enquanto método é muitas vezes considerado problemático e desafiador (O'BRIEN, 2009; ALVES et al., 2009). A abordagem de avaliação cognitiva é utilizada em um experimento com rastreamento ocular para determinar a complexidade dos erros na combinação linguística inglês-lituano dos mais fáceis aos mais difíceis, conforme visto pelos leitores do texto traduzido automaticamente. Os parâmetros testados (duração do

Belo Horizonte, v. 13, n. 2, mai.-ago. 2020 – ISSN 1983-3652
DOI:

olhar e número de fixações) demonstram que é necessário um esforço cognitivo diferente para processar diferentes tipos de erros em textos traduzidos de forma automática. Este trabalho almeja contribuir para outras pesquisas neste campo, pois fornece análise de avaliação de erros da produção de tradução automática.
PALAVRAS-CHAVE: tradução automática; abordagem de avaliação cognitiva; erro(s) de tradução; rastreamento ocular; aceitabilidade.

1 Introduction

Although eye tracking research methodology is not free of complexity and ambiguity, many studies in translation research rely on eye tracking as it has been long ago assumed and many a time proven that cognitive effort is reflected well by eye movement (ALVES et al., 2009; O'BRIEN, 2009; HVELPLUND, 2017). Eye tracking in studying reading for translation has also become a standard methodology (JAKOBSEN; JENSEN, 2008). Along with think-aloud protocols and pause measurement, eye tracking has been increasingly used to measure cognitive effort in machine translation research (MOORKENS, 2018). Kornacki (2019) has contributed to the field by trying to determine the applicability of eye tracking methodology in a computer-based translation classroom. However, such studies are still not numerous and take various research designs.

Studies on the cognitive effort of machine-translated output, many of which employ eye tracking methodology, are abounding (CARL et al., 2011, 2015; DAEMS et al., 2017; GONÇALVES, 2016; O'BRIEN, 2006, 2011; MOORKENS, 2018; SPECIA, 2011; CASTILHO, 2016). Some of them focus on the professional post-editors' (or translators') cognitive effort in processing machine translation (MT) output, quite often in comparison with novices in the translation field (ALVES et al., 2016; NITZKE, 2016). For example, Nitzke (2016) has compared semi-professional and professional translators in an experiment of translating from scratch, bilingual post-editing and monolingual post-editing by eye tracking and screen recording data. The study has explored the frequency of superficial mistakes (like grammar, spelling, etc.) and content mistakes in all three tasks and has found that the number of superficial mistakes is lower in the monolingual post-editing task in comparison with translation from scratch and bilingual post-editing, while content in case of the monolingual post-editing task is error-prone (NITZKE, 2016). Alves et al. (2016) have investigated the cognitive effort required by professional translators in post-editing tasks using interactive and non-interactive machine translation workbenches. The authors have found that less cognitive effort is required when an interactive machine translation workbench is used (ALVES et al., 2016).

Some other research studies employ eye tracking methodology to investigate the MT quality evaluation based on error analysis. Stymne et al. (2012) have conducted an MT error analysis in a task of identification and classification of MT errors by university students who tended to exert more effort in processing MT errors, as they were shown to have longer gaze times and greater fixation counts in comparison with an accurately translated text.

Our study employed eye tracking research methodology to evaluate MT output via an experiment of a reading task. The aim of the research was to determine the types of errors in machine translation output processed from English into Lithuanian that cause

Belo Horizonte, v. 13, n. 2, mai.-ago. 2020 – ISSN 1983-3652
DOI:

understanding problems to potential readers. The rationale behind choosing the English-Lithuanian pair for the experiment lies in the fact that Lithuanian is the so-called minor language, and machine translation systems for Lithuanian still need more extensive training to provide high quality. The implications obtained through such an experiment might be relevant for any language pairs, especially for small-scale languages. In this study, the cognitive evaluation approach was used in a reading task of machine-translated output. In order to achieve the aim, the following research questions were raised: Do errors in a machine-translated text require additional cognitive effort? What types of errors cause a longer gaze and a greater number of fixations? To what extent is the text acceptable to the readers of the machine-translated text?

The following hypotheses were raised:

a) The mean gaze time spent on the segments with errors is longer than on the segments without errors and the mean fixation count on the segments with errors is greater than on the segments without errors.
b) The mean gaze time spent and the mean fixation count are different on segments with different types of errors.
c) Overall acceptability of the raw machine-translated text obtained via a post-task survey correlates with the readers' gaze time spent on segments with errors.

The current work aims at contributing to other research in the Translation Studies field by providing analysis of error assessment of machine translation output in Lithuanian. To the best of our knowledge, such a study is the first attempt to use the cognitive evaluation approach in eye tracking as a supplementary technique to other existing ways of machine translation error analysis in the English-Lithuanian language pair.

2 Assessment of machine translation quality

Human assessment of machine translation quality has been considered significant despite the challenges and inconsistencies in the approach taken by scholars and assessors (GRAHAM, 2015). Different taxonomies for machine translation assessment have been proposed (see FLANAGAN, 1994; VILAR et al., 2006).

The first MT output quality assessment system was introduced by Flanagan (1994). Based on English to French machine-translation output, the author distinguished 21 major and minor categories of errors in spelling (misspelled word), not found word (word not in dictionary), accent (incorrect accent), capitalization (incorrect upper or lower case), elision (illegal elision or elision not made), verb inflexion (incorrectly formed verb or wrong tense), noun inflexion (incorrectly formed noun), other inflexion (incorrectly formed adjective or adverb), rearrangement (sentence elements ordered incorrectly), category (of nouns or verbs), pronoun (wrong, absent or unnecessary pronoun), article (wrong, absent or unnecessary article), preposition (wrong, absent or unnecessary preposition), negative form (negative particles not properly placed or absent), conjunction (failure to reconstruct constituents after conjunction or identify boundaries of joined units), agreement (incorrect subject-verb, noun-adjective, etc. agreement), clause boundary (failure to identify clause boundary or unnecessary clause boundary), word selection expression (word selection error or wrong translation of multi-word unit), relative pronoun (wrong or absent), case

(wrong case ending), and punctuation (wrong, absent or unnecessary) (FLANAGAN, 1994). The last three categories were added for the MT output assessment in the English–German language pair. However, Flanagan (1994) advocated for an individual category set to be developed for each language pair as certain error types are relevant only for certain languages and vice versa.

Another common taxonomy of machine translation output assessment was developed by Vilar et al. (2006) who classified errors of machine translation output into five fundamental categories: missing words, further subdivided into missing content or filler words; word order either at the word level or phrase level; incorrect words, subdivided into sense (wrong lexical choice or incorrect disambiguation), incorrect form, extra words, style, idioms; unknown words subdivided into unknown stems and unseen forms; and punctuation. Vilar et al.'s explicit error taxonomy was based on Chinese–English, Spanish–English and English–Spanish statistical MT systems. Since then, this hierarchical error classification of machine translation errors has been modified by other scholars to suit the needs of different language pairs or other purposes (see Popovic, 2018, for a detailed overview of error typologies).

Temnikova (2010, 2016) has regrouped Vilar et al.'s error taxonomy and ranked the errors from the easiest to the hardest to correct in order to reveal the cognitive effort that MT output correction requires. The easiest errors to correct were morphological, i.e., correct word, incorrect form. The medium errors to correct, requiring replacing or adding a word, were lexical, i.e., incorrect style, synonym, incorrect word, extra word, missing word, and idiomatic expression. The hardest errors were supposed to be syntactic, requiring understanding of the whole sentence, i.e., wrong punctuation, missing punctuation, word order at word level, and word order at phrase level. The authors claim that this approach is reliable, objective and valuable because it allows identifying and differentiating between the errors requiring more and less cognitive effort (TEMNIKOVA, 2016).

Although MT error taxonomies have been developed and modified many times by different scholars, there is a strong need to create an individualised error taxonomy for each different language pair, as advocated by Flanagan (1994). For the purposes of machine translation assessment in the English–Lithuanian language pair, there has been an attempt to present an adapted classification by Petkevičiūtė and Tamulynas (2011) who reinterpreted and regrouped categories, identified in the taxonomy by Hutchins and Somers (1992), into two broad types, namely linguistic (morphological and lexical) and systemic. In Petkevičiūtė and Tamulynas's (2011) terms, linguistic morphological errors were subdivided into errors in case, main verb form, number, person, gender, part of speech, negative verb, and missing verb. Lexical errors include untranslated phrase, untranslated word, hyphenated word, literal translation of phrases (added unnecessary words), contraction, polysemous word, pronoun, abbreviation, and proper name (ibid.). Systemic errors (or errors related to the source code) were considered as the ones lacking linguistic or logic explanation: errors in diacritics, an extra word that was not used in the original text, meaning of the word not in a dictionary, word translated into a different target language, missing word, capitalization) (ibid.). In their research, Petkevičiūtė and Tamulynas (2011) found that the most common errors were morphological errors, namely those of case, gender, main verb form, number and part of speech, as well as lexical errors, i.e., those of untranslated word and polysemy. Although the researchers interpreted the MT errors differently, they concluded that despite different classifications and

Belo Horizonte, v. 13, n. 2, mai.-ago. 2020 – ISSN 1983-3652
DOI:

interpretations, lexical and morphological errors would always be most important (PETKEVIČIŪTĖ; TAMULYNAS, 2011).

3 Acceptability

Human acceptability has been used as a criterion in evaluation of translation, including MT, quality. Many studies have followed Van Slype's definition of acceptability, which is "a subjective assessment of the extent to which a translation is acceptable to its final user" (1979). The author proposed to measure acceptability by way of survey questions (1979). Other prevailing definitions of acceptability also emphasise and focus on a degree that a text is acceptable (CHOMSKY, 1969), reader's attitude and tolerance towards the text (DE BEAUGRANDE; DRESSLER, 1981; ROTURIER, 2006), or usability, satisfaction and quality (CASTILHO, 2016). All these characteristics that define acceptability are of a subjective nature, thus making acceptability a vague notion.

According to Castilho et al. (2018), one of the measures of machine translation output quality is acceptability, which shows the reader's attitude towards texts in terms of correctness, cohesion and coherence. Even if the text contains errors, it may be acceptable as far as it serves the needs of the readers (CASTILHO, 2016). Acceptability is measured via usability (efficiency, effectiveness and cognitive effort), satisfaction (web survey, post-task satisfaction questionnaire and moderators' ratings) and quality (fluency, adequacy, syntax and grammar, and style in translated content and text easeability, readability, source content profiler score, and domain classification in source content) (CASTILHO 2016). Other authors have proposed to measure human acceptability of MT by way of survey questions (VAN SLYPE, 1979). For the purposes of this research, acceptability is understood as a notion combining satisfaction, usability and quality assumed by the readers of the text.

4 Research design and experiment

The study is based on the theoretical framework proposed by Petkevičiūtė and Tamulynas (2011), Temnikova (2010, 2016) and Vilar et al. (2006). Petkevičiūtė and Tamulynas's (2011) version of machine translation error typology was employed in the analysis to classify the errors found in the text machine-translated from English into Lithuanian. The idea proposed in Temnikova's research on cognitive evaluation approach towards MT output was employed to rank the machine translation errors from easiest to hardest and check whether the errors found in English–Lithuanian machine translation output may fit within Temnikova's original research approach (2010, 2016).

4.1 Participants and data collection

Eye movements of 14 subjects were tracked in a reading comprehension task with a text translated from English into Lithuanian by a freely available neural machine translation engine, namely Google Neural Machine Translation. The text used for the experiment was of a news type, published in the English language on a website of a

Belo Horizonte, v. 13, n. 2, mai.-ago. 2020 – ISSN 1983-3652
DOI:

worldwide known organization. As this text was a piece of news, the language was quite simple, non-sophisticated and meant for the general public. The subjects (13 women and 1 man) were native speakers of Lithuanian recruited for the experiment from the university staff. They all had a university diploma. Five respondents had a degree in languages, and the rest had a degree in other fields. The subjects gave consent to participate in the experiment on a voluntary basis. They were informed that the text they were reading was a translation with no specification on the translator – human or machine. The subjects were also told that they would have to answer reading comprehension questions afterwards and fill in a post-task questionnaire on the acceptability of the text.

Eye tracking was performed using a commercial non-invasive eye-tracking device, and an analysis software – gaze monitoring system – developed by information technologies specialists of Kaunas University of Technology (Lithuania) for the university research purposes (TURENKO et al., 2019). The experiment focused on areas of interest of the machine-translated text (segments with and without errors), which required a longer gaze time and a greater number of fixations. The raw machine-translated text contained 179 words. In total, 12 segments were marked as areas of interest by human error analysis performed before the experiment: 7 segments with linguistic morphological errors (4 segments with wrong case endings, 2 segments with errors in gender and 1 segment in the wrong use of a negative verb), 2 segments with lexical errors (1 segment with an untranslated phrase and 1 segment with a literal translation) and 3 segments with systemic errors (2 segments with an added unnecessary word and 1 segment with wrong capitalisation). There were in total 7 different types of errors (definitions are provided in the section on assessment of machine translation quality).

For the experiment, onscreen stimuli were presented as follows: screen resolution 1920 x 1080 pixels; font style Calibri (body); font size 22; multiple line spacing; maximum 110 characters including spaces or maximum 94 characters excluding spaces per line; maximum 16 words per line. There were 13 lines in the text fitted in one column so that no scrolling was required. Six lines contained one error, three lines contained two errors and three lines contained no errors. In two lines with two errors, the types of errors were different. In one line with two errors, the same type of errors was present. The distance between the errors in the lines with two errors was from three to six words. There was only one segment with an error at the very end of the line. See Figure 1 for a sample eye tracking map of one subject.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 13, n. 2, mai.-ago. 2020 – ISSN 1983-3652
DOI:

Figure 1: A sample eye tracking map of one subject.
Source: from the authors.

A retrospective survey was used as a complementary research method to test the end user acceptability of the machine translated text given that eye tracking as a research methodology is not free of subjectivity. Acceptability is here understood in terms of three criteria, namely, satisfaction, usability and quality. In this research, after the experimental reading task, the participants were given a post-task questionnaire consisting of two parts. Part 1 consisted of 6 statements given to the respondents on satisfaction, usability and quality of machine translated output (see Appendix 1). Statements 1, 2 and 3 in the post-task questionnaire were included to measure assumed user satisfaction with the translated text; statement 4 was asked to determine the usability of the text; and statements 5 and 6 were added to evaluate the quality. In total, in this part of the questionnaire, the subjects of the experiment could accumulate a maximum of 30 points. Part 2 of the post-task questionnaire included three open-type reading comprehension questions specifically related to the main idea and details of the text to find out the respondents' level of understanding of the text.

4.2 Data analysis

Specialised eye tracking software finds basic parameters (coordinates, time etc.) for each participant in a separate file. The data processing and visualization were performed using MS Excel 2016 and MS Access 2016. IBM SPSS Statistics 20 was used for the analysis of the collected data. Using MS Excel 2016, the data were aggregated by text segments, and the resulting files were merged into a single table, prepared for statistical analysis with IBM SPSS Statistics 20. MS Access 2016 was used to connect survey and experimental data (relations between tables and several queries were needed). Furthermore, in MS Access 2016, a new table was created, and the data were statistically analyzed using IBM SPSS Statistics 20.

IBM SPSS Statistics 20 was used for descriptive, comparative and relationship analysis. Descriptive statistics (the mean gaze time spent on the segments with errors and

without errors, percentages, etc.) were calculated for quantitative and qualitative data. Quantitative data (gaze time, gaze time percentage on a segment, fixation count on segments) were tested for the distribution normality using the Kolmogorov-Smirnov test. As the distribution of all variables was not normal, non-parametric tests were applied. Comparative analysis was performed using the Mann-Whitney test to find the existence of statistically significant differences between the gaze time on the segments with errors and the gaze time on the segments without errors, between the fixation count on the segments with errors and the segments without errors. The relationship analysis was carried out with the Spearman correlation to analyse fused survey and experimental data.

5 Results

The text used in the experiment was divided into 58 segments in total: 12 segments with one error in each segment and 46 segments containing no errors. The length of the segments in terms of the number of characters is presented in Figure 2. The segments with errors had a minimum of 11 characters (excluding spaces) and a maximum of 40 characters (excluding spaces), and a minimum of 3 words and a maximum of 5 words. The segments without errors had a minimum of 8 characters (excluding spaces) and a maximum of 40 characters (excluding spaces), and a minimum of 3 words and a maximum of 5 words.

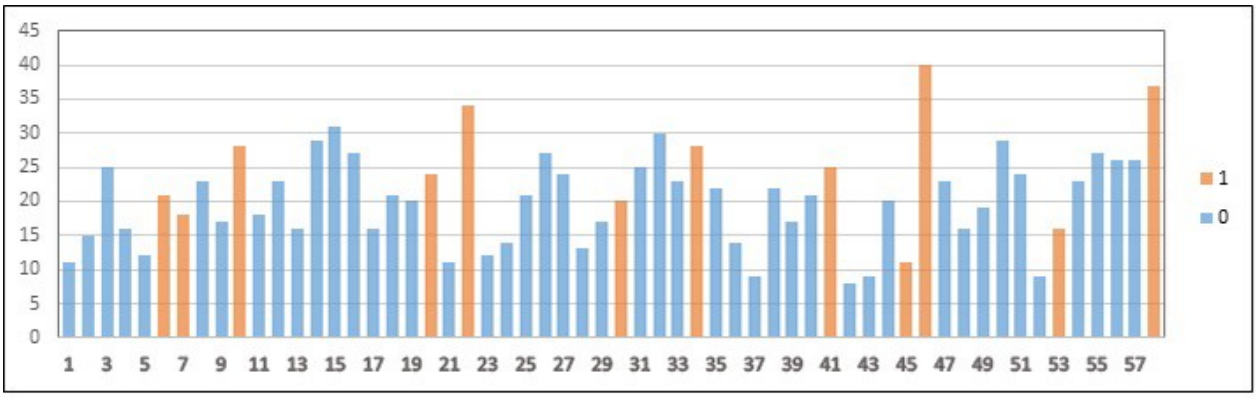

Figure 2: Number of characters in segments (excluding spaces). Blue bars indicate segments without errors; orange bars indicate segments with errors.
Source: from the authors.

The analysis of the findings demonstrated that the segments with machine translation errors required longer gaze times and more fixations than the segments with no errors (see Figures 3 and 4). The mean gaze time spent on the segments with errors was 1.83 ms in comparison with the mean time spent on the segments without errors, which was 1.48 ms (see Figure 3). The Mann-Whitney test demonstrated that there was a statistically significant difference between the gaze time on the segments with errors and the gaze time on the segments without errors ($Z = –3.305$, $p = 0.001$).

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 13, n. 2, mai.-ago. 2020 – ISSN 1983-3652
DOI:

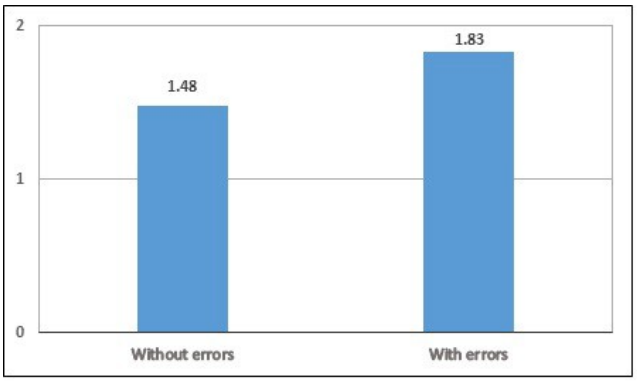

Figure 3: Mean gaze time of all participants spent on the segments without errors and the segments with errors.
Source: from the authors.

Overall, the mean fixation count on all segments with errors was 166 times; meanwhile, the mean fixation count on all segments without errors was 135 times (see Figure 4). The Mann-Whitney test demonstrated a statistically significant difference between the fixation count on the segments with errors and the segments without errors ($Z = –3.975$, $p < 0.001$).

Figure 4: Mean fixation count of all participants on the segments without errors and the segments with errors.
Source: from the authors.

Additionally, the mean gaze time and the fixation count were calculated for each segment with errors separately (see Figure 5). The longest gaze time was observed on the segment with an added unnecessary word error (mean gaze time 2.7 ms). The second longest gaze time was observed on the segment with a literal translation of a phrase (mean gaze time 2.56 ms), followed by another segment with an added unnecessary word (mean gaze time 2.37 ms) and a segment with a case error (mean gaze time 2.37 ms), followed by other segments with case errors (mean gaze time 2.25 ms, 2.25 ms).

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 13, n. 2, mai.-ago. 2020 – ISSN 1983-3652
DOI:

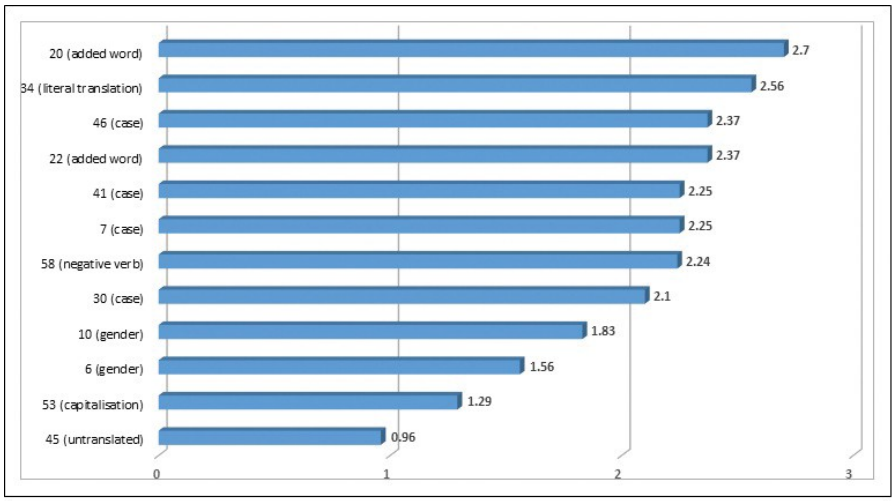

Figure 5: Mean gaze time for each segment with errors.
Source: from the authors.

The segments with errors that received the absolutely longest (mean gaze time 2.7 ms) and the third/fourth longest (mean gaze time 2.37 ms) gaze time contained added word errors. In both cases in this experiment, the translations resulted in a repetition of the same word. We may speculate that such segments received a longer gaze time because the subjects read two exactly similar words, which looks inexplicable, e.g., *clinical trials* was machine translated as *klinikiniai klinikiniai tyrimai* (back translation into English – *clinical clinical trials*). The second longest mean gaze time was observed on the error that contained literal translation (mean gaze time 2.56 ms). In Petkevičiūtė and Tamulynas's perspective, added unnecessary word errors and literal translation would be categorised as lexical errors. According to the findings of our research, it may be indicated that these errors required the highest cognitive effort. The findings of Temnikova's research showed this type of errors to be medium in terms of revealing the cognitive process that machine translation output requires.

The type of errors that overall required the second highest cognitive effort were case errors (third/fourth (mean gaze time 2.37 ms), fifth (mean gaze time 2.25 ms), sixth (mean gaze time 2.25 ms) and eighth (mean gaze time 2.1 ms) segments in terms of the longest gaze time spent), e.g., *other childhood vaccines* was translated as *kitų vaikų vakcinomis* (back translation into English – *vaccines of other children*). This was followed by gender errors (ninth (mean gaze time 1.83 ms) and tenth (mean gaze time 1.56 ms) segments in terms of the longest gaze time spent), e.g., *among the most extensively studied medical products* was translated as *viena iš plačiausiai tiriamų produktų* (back translation into English – *one of the most extensively studied products*). The numeral *viena* / *one* which is in the feminine form should be used here in concord with the masculine noun *produktų* / *products*, i.e., *vienas* / *produktų*. The seventh longest mean gaze time was observed on the segment with a negative verb error (mean gaze time 2.24 ms), e.g., *vaccine does not increase the risk of autism, or trigger autism* was translated as *vakcina nepadidina autizmo rizikos arba sukelia autizmą* (back translation into English – *vaccine does not increase the risk of autism, or triggers autism*). All these errors would be categorised as linguistic morphological errors in Petkevičiūtė and Tamulynas's perspective.

The segments that received the shortest gaze time contained capitalisation error

Belo Horizonte, v. 13, n. 2, mai.-ago. 2020 – ISSN 1983-3652
DOI:

(mean gaze time 1.29 ms) and an untranslated phrase error (mean gaze time 0.96 ms). It may be assumed that the latter segment received the shortest gaze time because it was in general the shortest segment with an error as it contained an abbreviation. However, presumably, an untranslated abbreviation should cause as much cognitive effort as other segments with errors or even more. In this case, we may speculate that the untranslated abbreviation (i.e., *CDC*) does not cause comprehension issues as the subjects of the experiment might have guessed from the context that the abbreviation referred to an organisation but were not interested in the exact name of the organisation. This brings to the fore one of the limitations of the research design. This segment has received a shorter gaze time (mean gaze time 0.96 ms) than the overall mean gaze time (mean gaze time 1.48 ms) spent on the segments with errors, just like the second shortest segment with errors, namely capitalisation (mean gaze time 1.29 ms), which may be the motive not to include such short segments with errors as areas of interest in future experiments. However, in order to make any assumptions or conclusions regarding errors involving abbreviations or capitalisation, more detailed and extensive research is needed. Whatsoever, the research findings proved that capitalisation and untranslated phrase errors, i.e., systemic errors in Petkevičiūtė and Tamulynas's perspective, required the lowest cognitive effort.

The results of the retrospective survey were analysed for three criteria composing acceptability, i.e., satisfaction, usability and quality. The Spearman correlation was used to determine if any relationship existed between the assumed readers' satisfaction and the time they spent on the segments with errors. The analysis of the data demonstrated a statistically significant correlation between the assumed readers' satisfaction and the time they spent on the segments with errors ($r = 0.642$, $p = 0.045$). There was no statistically significant correlation between the readers' assumed usability and quality scores and the times spent on reading segments with errors ($r = 0.043$, $p = 0.906$; $r = 0.055$, $p = 0.880$, respectively).

In part 2 of the post task-questionnaire, the respondents were asked to answer three text comprehension questions. The answers to the questions were checked later for the compliance with the information and specific details mentioned in the text. The questions were formulated to allow simple scoring. A correct answer was assigned 1 point, an incorrect answer was assigned 0 points and a partially correct answer was assigned 0.5 points. The respondents' answers to the text comprehension questions demonstrated that the level of understanding of the machine-translated text was relatively high. On the average, the mean combined score of correct answers for all subjects was 1.8 of 3 points (minimum 0 points and maximum 3 points).

6 Concluding remarks

In this paper, we calculated the mean gaze time and the mean fixation count for different types of errors and all errors present in the text separately, as shown by the cognitive effort spent on reading a machine-translated text from English into Lithuanian by fourteen uninformed subjects. The main aim was to determine by way of an eye tracking experiment the types of errors that cause understanding problems to potential readers in order to evaluate the machine translation output.

Belo Horizonte, v. 13, n. 2, mai.-ago. 2020 – ISSN 1983-3652
DOI:

The study was grounded on three hypotheses. The first hypothesis was related to the mean gaze time and the mean fixation count on the segments with errors in comparison with the segments without errors. The findings allowed us to confirm this hypothesis, namely that the mean gaze time spent on the segments with errors is longer than on the segments without errors and the mean fixation count on the segments with errors is greater than on the segments without errors.

The research also corroborated the second hypothesis, namely that the mean gaze time and the mean fixation count are different for segments with different types of errors: errors that receive the longest mean gaze time and the greatest fixation count are lexical errors; those that receive the medium mean gaze time and the medium mean fixation count are linguistic morphological errors and those that receive the shortest mean gaze time and the lowest mean fixation count are systemic errors. The results demonstrating the differentiation of errors by their complexity from the easiest to the hardest in our study (namely, systemic, morphological and lexical) contradict Temnikova's findings where the author observed that processing of lexical errors required medium cognitive effort. However, the findings of our study and Temnikova's research are only partially comparable because of different methodology, language pairs involved, etc.

The analysis of the results did not allow us to confirm the third hypothesis that overall acceptability of the raw machine-translated text obtained via a post-task survey correlates with the readers' gaze time spent on segments with errors. However, when the results for three different criteria, composing acceptability, namely satisfaction, usability and quality, were analyzed separately, a statistically significant correlation was observed between the users' satisfaction with the text and the time they spent on the segments with errors.

Based on these hypotheses and the findings, we may claim the following: errors in a machine-translated text require additional cognitive effort in comparison with error-free text segments. Lexical errors cause more cognitive effort than any other types of errors in English to Lithuanian machine translated text as demonstrated by the time and the fixation count on the segments with different types of errors. In order to determine the extent to which the machine-translated text is acceptable to the readers, further research is needed.

This study may provide a possibility to understand more deeply the readers' cognitive effort and the level of acceptability they exhibit towards machine-translated texts. Further and more extensive research is needed to replicate the data and investigate that the mean gaze time and fixation count are dependent on the error type. More data and evidence are needed in terms of the length of a machine-translated text in an experiment, the number of segments with errors and without errors, the number of types of different errors and more subjects with different background participating in the experiment in order to get more valid and reliable results.

References

ALVES, F.; SARTO SZPAK, K.; GONÇALVES, J. L.; SEKINO, K.; AQUINO, M.; ARAUJO E CASTRO, R.; KOGLIN, A.; DE LIMA FONSECA, N. B.; MESA-LAO, B. Investigating cognitive effort in post-editing: A relevance-theoretical approach. In: HANSEN-SCHIRRA, S. and GRUCZA, S. (eds.). *Eyetracking and applied linguistics*, Berlin: language science

press, 2016. p. 109-142.

ALVES, F.; PAGANO, A.; DA SILVA, I. A new window on translators' cognitive activity: methodological issues in the combined use of eye tracking, key logging and retrospective protocols. In: MEES, I. M.; ALVES, F.; GÖPFERICH, S. (eds.). *Methodology, Technology and Innovation in Translation Process Research. Copenhagen*: Samfundslitteratur, 2009. p. 267-292.

CARL, M.; DRAGSTED, B.; ELMING, J.; HARDT, D.; LYKKE JAKOBSEN, A. The process of post-editing: a pilot study. In: *Proceedings of the 8th International Natural Language Processing and Cognitive Science Workshop*, p. 131-142, 2011.

CARL, M.; GUTERMUTH, S.; HANSEN-SCHIRRA, S. Post-editing machine translation: Efficiency, strategies, and revision processes in professional translation settings. Psycholinguistic and Cognitive Inquiries into Translation and Interpreting. In: FERREIRA, A.; SCHWIETER, J. W. (eds.). *Psycholinguistic and cognitive inquiries into translation and interpreting*. Amsterdam: John Benjamins, 2015. p. 145-174.

CASTILHO, S. *Measuring Acceptability of Machine Translated Enterprise Content* (PhD thesis). Dublin: Dublin City University, 2016.

CASTILHO, S.; DOHERTY, S.; GASPARI, F.; MOORKENS, J. Approaches to Human and Machine Translation Quality Assessment. In: MOORKENS J.; CASTILHO S.; GASPARI F.; DOHERTY S. *Translation Quality Assessment.* Machine Translation: Technologies and Applications, 2018. p. 9-38.

CHOMSKY, N. *Aspects of the theory of syntax*. Cambridge: MIT Press, 1969.

DAEMS, J.; VANDEPITTE, S.; HARTSUIKER, R. J.; MACKEN, L. Identifying the machine translation Error types with the greatest impact on post-editing effort. In: *Frontiers in Psychology*, v. 8, 2017. p. 1-15.

DE BEAUGRANDE, R.; DRESSLER, W. *Introduction to text linguistics*. London, New York, 1981.

FLANAGAN, A. M. Error classification for MT evaluation. In: *Proceedings of the first Conference of the Association for Machine Translation in the Americas,* 1994. p. 65-72.

GONÇALVES, J. L. Investigating Saccades as an index of cognitive effort in post-editing and translation. In: *Proceedings of EST Congress.* 2016, Aarhus, Denmark, 2016.

GRAHAM, Y. Improving Evaluation of Machine Translation Quality Estimation. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2015. p. 1804-1813.

HVELPLUND, K. T. Four fundamental types of reading during translation. In: JAKOBSEN,

A. L.; MESA-LAO, B. (eds.), *Translation in Transition*: Between Cognition, Computing and Technology. Amsterdam / Philadelphia: John Benjamins Publishing Company, 2017. p. 55-77.

HUTCHINS, J.; SOMERS, H. *An introduction to machine translation*. London: Academic Press, 1992.

JAKOBSEN, A. L.; JENSEN K. T. H. Eye movement behaviour across four different types of reading task. In: GÖPFERICH, S.; JAKOBSEN, A. L.; MEES, I. M. (eds.). *Looking at eyes*: eye-tracking studies of reading and translation processing, 2008. p. 78-98.

KORNACKI, M. The application of eye-tracking in translator training. *inTRAlinea*: New Insights into Translator Training, 2019.

MOORKENS, J. Eye tracking as a measure of cognitive effort for post-editing of machine translation. In: WALKER, C.; FEDERICI, F. M. (eds.). *Eye tracking and multidisciplinary studies on translation*, 2018. p. 55-69.

NITZKE, J. Monolingual post-editing: An exploratory study on research behavior and target text quality. In: HANSEN-SCHIRRA, S.; GRUCZA, S. (eds.). *Eyetracking and applied linguistics*, Berlin: Language Science Press, 2016. p. 83-109.

O'BRIEN, SH. Towards predicting post-editing productivity. *Machine Translation*, v. 25, n. 3, p. 197-215, 2011.

O'BRIEN, SH. Eye tracking in translation-process research: methodological challenges and solutions. In: MEES, I. M.; ALVES, F.; GÖPFERICH, S. (eds.). *Methodology, Technology and Innovation in Translation Process Research. Copenhagen*: Samfundslitteratur, 2009. p. 251-266.

O'BRIEN, SH. *Machine-translatability and post-editing effort*: An empirical study using translog and choice network analysis. Dublin: Dublin City University, 2006.

PETKEVIČIŪTĖ, I.; TAMULYNAS, B. Kompiuterinis vertimas į lietuvių kalbą: alternatyvos ir jų lingvistinis vertinimas. *Studies about Languages*, v. 18, Kaunas: Technologija, p. 38-45, 2011.

POPOVIC, M. Error classification and analysis for machine translation quality assessment. In: MOORKENS, J.; CASTILHO, S.; GASPARI, F.; DOHERTY, S. (eds.). *Translation quality assessment:* from principles to practice, Springer international publishing AG, 2018. p. 129-158.

ROTURIER, J. *An investigation into the impact of controlled English rules on the comprehensibility, usefulness and acceptability of machine-translated technical documentation for French and German users* (PhD thesis). Dublin: Dublin City University, 2006.

SPECIA, L. Exploiting objective annotations for measuring translation post-editing effort. In: FORCADA, M. L.; DEPRAETERE H.; VANDEGHINSTE V. (eds.). *Proceedings of the 15th Conference of the European Association for Machine Translation*, 2011. p. 28-37.

STYMNE, S.; DANIELSSON, H.; BREMIN, S.; HU, H.; KARLSSON, J.; PRYTZ LILLKULL, A.; WESTER, M. Eye Tracking as a Tool for Machine Translation Error Analysis. In: CALZOLARI, N.; CHOUKRI, K.; DECLERCK, T.; UĞUR DOĞAN, M.; MAEGAARD, B.; MARIANI, J.; MORENO, A.; ODIJK, J.; PIPERIDIS, S. (eds.). *Proceedings of the 8th international conference on language resources and evaluation*, 2012. p. 1121-1126.

TEMNIKOVA, I. Cognitive evaluation approach for a controlled language post-editing experiment. In: *Proceedings of the 7th international conference on language resources and evaluation*, 2010. p. 3485-3490.

TEMNIKOVA, I.; ZAGHOUANI W.; VOGEL, S.; HABASH, N. Applying the Cognitive Machine Translation Evaluation Approach to Arabic. In: *Proceedings of the International Conference on Language Resources and Evaluation*, 2016. p. 3644-3651.

TURENKO, V.; BALTULIONIS, S.; VASILJEVAS, M.; DAMAŠEVIČIUS, R. Analysing program source code reading skills with eye tracking technology, Information Technology. In: *Proceedings of IVUS'2019*, Vytautas Magnus University, 2019. p. 50-54.

VAN SLYPE, G. *Critical study of methods for evaluating the quality of machine translation*. Brussels: Buerau Marcel van Dijk, 1979.

VILAR, D.; XU, J.; D'HARO, L. F.; NEY, H. Error analysis of statistical machine translation output. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006. p. 697-702.

Recebido em dia 12 de maio de 2020.
Aprovado em dia 07 de julho de 2020.

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 13, n. 2, mai.-ago. 2020 – ISSN 1983-3652
DOI:

APPENDIX 1. Statements* evaluated on a 5-point Likert scale** in order to check the readers' satisfaction, usability and quality of machine translated output

1. The main idea of translated text was easy to understand.
2. The details of translated text were easy to understand.
3. The language (grammar, lexis) was easy to understand.
4. The translation is suitable for publication.
5. The quality of text is excellent.
6. The sentences in the text sound natural.

* The statements were provided to the subjects as a post-task survey in their native, i.e., Lithuanian, language.

** 1 – disagree, 2 – somewhat disagree, 3 – neither disagree, nor agree, 4 – somewhat agree, 5 – agree.