A Reception Study of Machine Translated Subtitles for MOOCs

Journal:	*Perspectives: Studies in Translation Theory and Practice*
Manuscript ID	PST-1110.R2
Manuscript Type:	Special Issue Paper
Keywords:	Machine Translation, Post-editing, Subtitles, Reception, MOOC
Abstract:	As access has grown to online courses in the form of MOOCs (Massive Open Online Courses), the language barrier has become an important issue for the users worldwide. Machine translation (MT) appears to offer an alternative or complementary solution to existing forms of MOOC translation. Very little attention has been paid, however, to the use and utility of MT for MOOC content. The main goal of this research is to test the impact machine-translated subtitles have on Chinese viewers' reception of MOOC content. We are interested in whether there is any difference between viewers' reception of raw machine translated subtitles as opposed to fully post-edited machine translated (PEMT) subtitles and human translated (HT) subtitles. Based on an eye-tracking experiment conducted at two Chinese universities and survey methods, we show that participants who were offered full PEMT subtitles scored better overall on our reception metrics than those who were offered raw MT subtitles. HT subtitles, on the other hand, did not necessarily lead to better reception as expected; in contrast, the participants who were offered HT subtitles performed the worst in some of our reception metrics.

A Reception Study of Machine Translated Subtitles for MOOCs

Abstract: As access has grown to online courses in the form of MOOCs (Massive Open Online Courses), the language barrier has become an important issue for the users worldwide. Machine translation (MT) appears to offer an alternative or complementary solution to existing forms of MOOC translation. Very little attention has been paid, however, to the use and utility of MT for MOOC content. The main goal of this research is to test the impact machine-translated subtitles have on Chinese viewers' reception of MOOC content. We are interested in whether there is any difference between viewers' reception of raw machine translated subtitles as opposed to fully post-edited machine translated (PEMT) subtitles and human translated (HT) subtitles. Based on an eye-tracking experiment conducted at two Chinese universities and survey methods, we show that participants who were offered full PEMT subtitles scored better overall on our reception metrics than those who were offered raw MT subtitles. HT subtitles, on the other hand, did not necessarily lead to better reception as expected; in contrast, the participants who were offered HT subtitles performed the worst in some of our reception metrics.

Keywords: machine translation; post-editing; subtitles; reception; MOOC

1. Introduction

Massive Open Online Courses, or 'MOOCs', have become an important pillar in China's thriving e-learning industry: according to its Ministry of Education, China is now the world's largest MOOC provider in terms of volume (MOE, 2018). But despite the growth in Chinese-language MOOCs, a considerable number of MOOCs accessed in China are in other languages, due to collaboration between Chinese and foreign MOOC platforms or universities. For example, one of the leading MOOC platforms in China, 学堂在线 (http://www.xuetangx.com/), is in partnership with edX and provides MOOCs from Stanford University, Queensland University, and the University of California, Berkeley, among others. The use of MOOCs in languages other than

Chinese can present problems however. In one survey on MOOC usage in China that attracted some 3,300 responses (MOOC 学院, 2014), language was cited as an important barrier to learning via MOOCs: of the more than 2,440 respondents who answered that they had tried MOOCs before, 47% claimed that the obstacle that stopped them learning via MOOCs was 'language'. For the respondents who answered they had not tried MOOCs before, 17.5% gave the 'language barrier' as the reason. There appears therefore to be scope for translation in the popularisation of MOOCs in China. Volunteer groups clearly have a role to play in such translation (Beaven et al. 2013), and but there is still a large unmet demand for translated MOOCs. Given limited availability of human translators and budget for translating content, Machine Translation (MT) could be part of the solution.

Over the past decade or so, several projects have addressed subtitling by MT, including MovRat (Armstrong et al., 2006; Flanagan, 2009), Volk's 'Stockholm system' (Volk, 2008), SUMAT (Etchegoyhen et al., 2014) and DialogueMT (Wang et al., 2016). Research in the area has been supported by the increasing availability of large parallel corpora of subtitles (e.g. Lison & Tiedemann, 2016), However, these projects mostly focus on films and TV shows, and subtitling MOOCs by MT is a fledging field. That said, the recently established TraMOOC project (Kordoni et al., 2016) focuses on the MT of subtitles for MOOCs. The project has established an open-source online neural MT platform (http://www.translexy.com/) that can automatically translate English-language MOOCs into 11 languages including Chinese. Despite such initiatives, MT still has an image problem and is not always trusted by end users. For example, administrators at OpenSubtitles.org have decided to get rid of all machine-translated subtitles, to avoid an accumulation of 'mass trash'[1]. To improve the quality of

MT output, and to establish trust among human subtitlers and end-users, post-editing could be a solution.

Post-editing means to 'edit and correct machine translation output' (ISO, 2017), Several sources (Plitt & Masselot, 2010; Aranberri et al., 2014) have reported productivity increases when using post-editing compared to traditional human translation. Castilho et al. (2014) further show that post-editing significantly increases the usability of machine-translated (online help) text. Their work sheds light on how end-users engage with raw and post-edited machine translated text, which is also the focus of our research. More specifically, the research presented in the current article investigates Chinese end users' reception of MOOC subtitles that have been translated from English into Chinese in three conditions: human translation (HT); raw, i.e., un-edited, MT (RAW); and post-edited MT (PEMT). The research revolves around one question: is there a difference in reception between participants who are offered raw MT subtitles and those who are offered full PEMT subtitles or HT subtitles?

The remainder of this paper is organized as follows: Section 2 outlines the methodology and briefly discusses our research setting. Section 3 focuses on data analysis and Section 4 on further discussion. Section 5 presents our conclusions and suggestions for future work.

2. Methodology

This research uses a mixed-methods approach that combines eye-tracking and questionnaire methods, following previous studies such as Caffrey's (2009) investigation of abusive subtitling of TV anime, Perego et al.'s (2010) analysis of the cognitive effectiveness of subtitle processing, Doherty's (2012) research on the effects of controlled language on the reading and comprehension of machine translated texts,

Secară's (2017) treatment of creative spellings in subtitling, Castilho's (2016) study on measuring acceptability of machine translated enterprise content, and Filizzola's (2016) work on Italians' perception and reception of British stand-up comedy humour, to name but a few.

We use eye-tracking to elicit quantitative data on the cognitive processing of translated subtitles and questionnaires to elicit qualitative, perception data. We proceed from the assumption that 'eye-tracking data can be interpreted as correlates of on-going cognitive processing of source and/or target texts' (Alves et al., 2012, p. 6). Questionnaire data on the other hand, reveal the traits and subjective experiences of participants, which helps us to discover any links between their perception and reception of subtitles, as operationalized below.

Consistent with the prior relevant studies, we believe that the combination of eye-tracking and questionnaire can provide a robust set of methods to investigate the perception and reception of subtitled AV content. To the best of our knowledge, this mixed-methods approach has not been adopted in any other research on machine translated subtitles of MOOC content to date.

2.1. Measurements and hypotheses

The concept 'reception' is operationalized in this research using Gambier's (2009) model, which is based on the three R's: response, reaction and repercussion, core concepts that coincide with those targeted in our experiment. Table 1 shows how we have interpreted Gambier's model for the purposes of our research[2]:

Table 1. Reception Model and Associated Measurement Tools (adapted from Gambier 2009)

Response refers to the initial physical response of a viewer to an audio-visual stimulus, in this case the subtitle and the rest of the MOOC image. In our study it was measured using glance count, glance duration, and fixation count[3], which indicate where the viewer's attention is directed. These data and other eye-tracking data were captured using an SMI REDn Scientific eye-tracker, which collected data at a rate of 60 Hz.

Reaction involves the cognitive follow-on from initial response, and is concerned with how much effort is involved in processing the initial stimulus and what is understood by the viewer. It was measured partly through average fixation duration[4], a typical effort indicator in eye tracking, and partly through testing viewers' comprehension of the MOOC content using specific questionnaire items.

Repercussion refers to attitudinal and sociocultural dimensions of AVT consumption. It was also captured using targeted questionnaire items.

In the light of the main research question mentioned at the end of Section 1, our main hypothesis is as follows: participants who are offered full PEMT subtitles and HT subtitles will score higher on our reception metrics compared with those who are offered raw (unedited) MT subtitles. This hypothesis is based on the assumption that raw MT output can be faulty, which, would have a negative effect on user's reception of the MOOC. It has to be noted that the main focus of this study is to compare raw machine translation with post-edited translation and human translation. However, a comparison between post-edited translation and human translation was also conducted

although we assumed that there would be fewer quality differences between these two conditions.

Based on the reception model and the main hypothesis, several sub-hypotheses have been derived as follows:

Regarding 'response':

Hypothesis 1a: More attention is allocated to the subtitle when raw MT subtitles are displayed than when full PEMT subtitles or HT subtitles are displayed.

This hypothesis was tested by measuring the glance count (or number of 'visits') in the subtitle area of interest ('AOI_SUB') of each video. The higher the glance count in AOI_SUB, the more attention participants were deemed to give to the subtitles.

Hypothesis 1b: More attention is allocated to the image area ('AOI_IMA') when full PEMT subtitles or HT subtitles are displayed than when raw MT subtitles are displayed.

To test this hypothesis, the glance count in the image AOI for each recording was captured. The higher the glance count in AOI_IMA, the more attention participants were deemed to give to the image.

Regarding 'reaction':

Hypothesis 2: The level of comprehension is higher with full PEMT subtitles and HT subtitles than with raw MT.

This hypothesis was tested using 13 comprehension questions, included in the post-task questionnaire described below.

Hypothesis 3: Average fixation duration is shorter when full PEMT subtitles or HT subtitles are displayed than when raw subtitles are displayed.

The average fixation duration for both image and subtitle areas was compared between the three conditions (HT, RAW and PEMT).

Regarding 'repercussion':

Hypothesis 4: Attitudes are more positive among participants shown full PEMT subtitles or HT subtitles.

This hypothesis was tested using 14 attitude statements, contained in the second part of the post-task questionnaire. The statements were evaluated by respondents using a five-point Likert scale. The data collected were coded into numerical values ranging from 1 for 'strongly disagree' to 5 for 'strongly agree'. The higher the score, the more positive the attitude to machine translation. Below are two examples of the statements (in English translation):

(1) The subtitles allow me to fully understand the contents of the MOOC.

☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

(2) The subtitles are useful to me.

☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

The majority of the statements in this part were specifically about the subtitles that had just been presented and only two (Q22 and Q23) were general statements on MT.

2.2. Research setting

According to a 2016 Chinese MOOC Industry Research White Paper (HRC, 2016), in China, MOOCs are mostly popular among university students between 18 and 25 years old. According to EF[5], a well-known international English education company, the English Proficiency Index (2017) for China is 52.45 out of 100, which indicates that Chinese people have a low proficiency in English, broadly speaking. To ensure representative samples, it was therefore decided to recruit Chinese undergraduates with low proficiency in English as participants for this research. We thus recruited 66 participants from two mid-sized universities in Anhui Province, China. Prior to recruitment, ethics approval was granted from the Research Ethics Committee of Dublin City University, in line with the requirements of all three universities. Participants were divided into three groups to compare the cases of users receiving raw MT subtitles with users receiving full PEMT or HT subtitles.

The video selected for this research was entitled 'What is physical activity?'. It appeared in the MOOC 'Sit Less, Get Active' on Coursera and was used with the latter's approval. The video lasted just under seven minutes (6'59") and was composed of vivid talks and colourful images. Speakers ranged from children to older people with different accents. All of these features made the video interesting and challenging for viewers with limited English. None of the participants had taken this MOOC before.

The MT tool used to translate the subtitles was Google Translate (Neural MT). While customized engines have since been created for the MT of MOOC content (for example, in the TraMOOC project as outlined above), these engines were not sufficiently developed at the time our experiment was conducted, and Google Neural MT outperformed competing systems in the initial tests that we ran. In the interest of giving MT a 'fair trial', we decided to use the best available system at the time. We were also aware that users who need free, instant translation are likely to turn to online systems such as Google Translate, which boasts some 500 million monthly users and translates over 140 billion words a day (Schuster et al., 2016). Full post-editing of the machine-translated subtitles was conducted by one of the authors, who is also an experienced EN-ZH translator. Human-translated subtitles were produced by a Chinese native speaker who is an experienced high school English teacher. The number of lines of subtitles was 135, 138 and 141, for raw MT, PEMT and HT respectively. Using the online Tilde BLEU score calculator[6] and tercom (Snover et al., 2006), and taking the human-translated subtitles as a reference in both cases, the BLEU score for the raw machine translated subtitles was found to be 42.05%, and the HTER for the post-edited subtitles to be 19.69%, indicating that the quality of the raw MT subtitles is relatively good.

The experiment involved three steps. In Step 1, participants completed an online pre-task questionnaire, conducted through 问卷星 (www.wjx.cn) and designed to collect demographic information, and an online English test to measure English proficiency. In Step 2, participants were asked to watch the MOOC video on a laptop fitted with an eye-tracker. In Step 3, participants completed a post-task questionnaire with two parts: comprehension testing and attitude survey. Most participants completed Step 1 using their mobile phones in their dorm or elsewhere. Steps 2 and 3 were carried

out in a dedicated room on campus under supervision of one of the researchers. Both pre-task and post-task questionnaires were administered in Chinese. Participants were not aware of whether the subtitles they saw were RAW, PEMT or HT.

3. Data analysis

3.1. Demographic Profile

Results show that the most typical respondent profile in our research is that of a '20-year-old' (46%), 'Year 3 undergraduate' (65%), with a 'Chemistry background' (73%). As many as 92% of participants had used MT before: 67% of them started using it since commencing university, and more than half of them (34) used MT every month. 84% of participants believed the quality of MT was 'not bad' or 'good'. However, only 9% believed MT could fully transfer the meaning of the source language text.

3.2. Online English test

The online English proficiency test[7] follows the Common European Framework of Reference for Languages[8]. Among the 61 participants who completed the full experiment, P26, P34, P36 and P38 from Group RAW, and P49, P51 and P54 from Group HT had Level C in the English test. Rather than eliminate the data for these seven participants, it was decided to scrutinize their comprehension score first to see if they scored higher than other participants: if these participants had a varied performance then it may not be necessary to remove them from the data analysis. The comprehension score of the seven participants is presented in Table 2. Compared to their group means (Group RAW = 8.55, Group HT = 9.47), it can be seen that their comprehension scores vary considerably. A high English score is not necessarily related to a high

comprehension score. Therefore, it was decided that the experimental data of the seven participants would not be removed from the data analysis in this paper.

Table 2. Comprehension testing score of the participants reporting Level C competence.

3.3. Post-task questionnaire

As already indicated, the post-task questionnaire had two parts: comprehension testing and attitude survey (see Appendix).

3.3.1. Comprehension testing

There were 13 questions in this part. Correct answers were assigned a score of 1; wrong answers a score of zero. Table 3 presents the comprehension testing score per group. We can see that the highest score and the mean of Group PE is higher than the other two groups. Regarding the mode of each group, for Group PE, six participants scored 10 and another six scored 11; for Group RAW, seven participants scored 8; and for Group HT, three participants scored 9, three scored 10, and another three scored 11.

Table 3. Comprehension testing score per group.

Apparently, Group PE performed the best in the comprehension test. To validate this and to test for statistical significance, a one-way ANOVA (Analysis of Variance) was carried out, which showed a statistically significant difference between the comprehension testing score of the three groups.

Since a significant ANOVA does not imply where that difference lies in the data, the LSD (least significant difference) test needs to be conducted to compare group means. The formula for LSD is as follows:

$$LSD = t\sqrt{MSw\ (\frac{1}{N1} + \frac{1}{N2})}$$

Where:

t = critical value from the t-distribution table

MSw = mean square within

N = number of scores used to calculate the means

In this case, the t-critical value for (α = 0.05, dfw = 56) is 2.003[9]. Here, the means from Group PE and Group RAW are compared first. When the given values are inserted into the LSD formula, we obtain the following result:

$$\text{LSD} = 2.003\sqrt{2.304\ (\frac{1}{22} + \frac{1}{22})} = 0.917$$

Our ANOVA found that the mean of Group PE is 9.91, and the mean of Group RAW is 8.55. Hence, the absolute value of the difference between these two means is 1.36. This value is greater than 0.917, indicating that there is a statistically significant finding between Group PE and Group RAW. Using the same method, no significant difference was found between Group PE and Group HT (0.44 < LSD = 1.017), or Group RAW and Group HT (0.92 < 1.017). In summary, the differences between the three groups are as follows:

Comparison 1: Group PE vs. Group RAW: significantly different

Comparison 2: Group RAW vs. Group HT: not significantly different

Comparison 3: Group HT vs. Group PE: not significantly different

It is already assumed that the quality of PEMT subtitles and HT subtitles is high, and that of raw MT subtitles is lower, though not substantially (See Section 2.1), which

means the results of Comparisons 1 and 3 are as expected. However, Comparison 2 turns out to be contrary to our expectation. As mentioned in Section 2.2, the human translation was conducted by an experienced English teacher. The HT subtitles were proofread by one of the researchers. Hence, it can be assured that the quality of the HT subtitles is high. In addition, Section 2.2 explained that taking HT subtitles as reference, the BLEU score for the raw MT subtitles was 42.05%, indicating that there is a difference between the two sets of subtitles. Therefore, reasons for the result of Comparison 2 remain unknown. Section 3.4 investigates this issue further, from the perspective of eye-tracking data.

3.3.2. Attitude Survey

This part consists of 14 five-point Likert scale statements as discussed in Section 2.1. Table 4 shows the distribution of different answers for each group in percentage form (absolute values are given in parentheses). For example, if all the participants in Group PE chose 'Strongly agree' for all the 14 statements, then this scale would have been chosen 336 times (24×14). However, it was chosen 114 times by Group PE because they did not 'strongly agree' with all the statements.

Table 4 shows that for all groups, most answers go to 'Agree', fewest answers go to 'Disagree' and 'Strongly disagree', indicating that participants' attitudes to the subtitles are mostly positive. In respect of the option 'Strongly agree', the number for Group PE is higher than that of the other two (33% vs. 17% vs. 27%). Though Group RAW has a slightly higher number in 'Agree' than Group PE, taking all the numbers

into consideration, the attitude of Group PE was overall better than the other two groups.

Table 4. Attitude survey results per group.

For each statement, the percentage of agreements (including 'strongly agree' and 'agree') per group was calculated. Results show that Group PE outperformed the other two groups on all statements except numbers 20, 25 and 27, indicating that the attitude of Group PE was overall better than others. In regard to Statement 20, it is surprising that the percentage of agreements for Group PE was the lowest (45.83%) and that for Group RAW was the highest (59.09%). In other words, Group RAW enjoyed reading their subtitles more than the other two groups. In regard to Statement 25, 'I could understand the subtitles if I could call someone for help if I got stuck', the percentage of agreements for Group RAW (77.27%) was higher than that for Group PE (66.67%) and Group HT (57.14%). The results for the two statements go against the conventional wisdom that raw MT would be more challenging and less enjoyable. Regarding Statement 27, 'I could understand the subtitles if I had just the built-in help facility for assistance', the percentage of agreements for Group HT was the highest (71.43%) while that for Group RAW was the lowest (68.18%). Looking at the number of times each of the three groups had the lowest mean score for a particular statement, it emerges that Group PE had the lowest score on one occasion (S20), Group RAW on four occasions (S16, S17, S19 and S27), and Group HT on nine occasions (S14, S15, S18, S21, S22, S23, S24 and S26). Similarly to what Comparison 2 shows in Section 3.3.1, this result is undoubtedly against expectation, mostly because HT subtitles received the most negative feedback while they are normally expected to have the highest quality. The fact that Group HT had the fewest participants may influence this result. However, this point requires more investigation when analysing the eye-tracking data for Group HT.

3.4. Eye-tracking analysis

Of the 61 participants who completed the full experiment, eye-tracking data for 39 reached the 70% tracking ratio threshold established for this research. In other words, the eye-tracker failed to track just over one third of the participants. As Blignaut and Wium (2014) suggest however, the trackability of Asian participants' eye-tracking data can be expected to be lower than for other ethnicities, and this fact needs to be taken into account when designing eye-tracking research. It is assumed that this has to do with the darkness of the Asian eye and eye tracking manufacturers have tried to compensate by having dark/bright pupil tracking, as described by Tobii[10]. Despite these advances, there may still be issues with tracking darker eyes. After eliminating those who had invalid eye-tracking data, there were 12 participants in Group PE, 14 participants in Group RAW and 13 participants in Group HT.

According to the data exported from the SMI eye-tracker, the export end trial time (visible time) was 419,000 milliseconds. Two AOIs were defined on the video for analysis: subtitle area (AOI_SUB) and image area (AOI_IMA). The coverage of AOI_SUB was 20.2% while that of AOI_IMA was 74.9%.

We expect that less demand on attention is required if the subtitles are easy to process. Hypotheses 1a and 1b assume that the better the quality of subtitles, the more attention participants would pay to the image area, and the less they would pay - or have to pay - to the subtitle area.

Hypothesis 1a: More attention is allocated to the subtitle area (AOI_SUB) when raw MT subtitles are displayed than when full PEMT subtitles or HT subtitles are displayed.

Hypothesis 1b: More attention is allocated to the image area (AOI_IMA) when full PEMT subtitles or HT subtitles are displayed than when raw MT subtitles are displayed.

The two hypotheses are measured by the glance count in the two AOIs of each video. The higher the number of glances in the AOI, the more attention participants are deemed to give to it. Thus, four sub-hypotheses are proposed as follows:

Hypothesis 1a.1: Group RAW > Group PE & Group HT
Hypothesis 1a.2: Group PE > Group HT
Hypothesis 1b.1: Group RAW < Group PE & Group HT
Hypothesis 1b.2: Group PE < Group HT

Table 5. Glance count in both AOIs of each group.

According to the means in Table 5, Group RAW had more glances in AOI_SUB (mean = 147.54) than Group PE (mean = 142.36) and Group HT (mean = 125.61), thereby supporting Hypothesis 1a. However, Group RAW (140.46) also had more glances in AOI_IMA than Group PE (139.82) and Group HT (133.42). (P14, P29 and P52 are outliers, and have been removed from analysis.) Thus, for the group that viewed the raw MT subtitles, more glances were given to both AOIs, which contradicts Hypothesis 1b.

One-way ANOVA[11] found no significant differences between the glance counts for three group pairs: Group RAW vs. Group PE, Group RAW vs. Group HT, and Group PE vs. Group HT.

The additional measures of fixation count and glance duration for both AOIs for each group are shown in Tables 6 and 7. Similarly to glance count, four sub-hypotheses for fixation count are proposed as follows:

Hypothesis 1b.3.1: AOI_SUB: Group RAW > Group PE & Group HT

Hypothesis 1b.3.2: AOI_SUB: Group PE > Group HT

Hypothesis 1b.4.1: AOI_IMA: Group RAW < Group PE & Group HT

Hypothesis 1b.4.2: AOI_IMA: Group PE < Group HT

Table 6. Fixation count for both AOIs for each group.

Table 6 shows that regarding fixation count, Group PE had the highest value in AOI_SUB (753.55), which contradicts Hypothesis 1b.3.1. Group PE had a higher value than Group HT (517.08), which supports Hypothesis 1b.3.2. Regarding the image area, Group RAW had the lowest value (351.46), which supports Hypothesis 1b.4.1. However, Group PE (414.64) had a slightly higher value than Group HT (413.42). The only two statistically significant differences for the fixation count measure were for AOI_SUB RAW vs HT and PE vs HT.

Table 7. Glance Duration [s] in both AOIs of each group.

As for glance duration, four hypotheses are proposed as follows:

Hypothesis 1b.5.1: AOI_SUB: Group RAW > Group PE & Group HT

Hypothesis 1b.5.2: AOI_SUB: Group PE > Group HT

Hypothesis 1b.6.1: AOI_IMA: Group RAW < Group PE & Group HT

Hypothesis 1b.6.2: AOI_IMA: Group PE < Group HT

Table 7 shows that regarding mean glance duration, Group RAW had the highest value in AOI_SUB (233.76), and the lowest value in AOI_IMA (133.00), which supports Hypothesis 1b.3 and Hypothesis 1b.4. The only statistically significant

difference for the glance duration measure lies in AOI_SUB RAW vs HT, with RAW subtitles having the highest glance duration on average, and RAW image having the lowest.

Hypothesis 3: Average fixation duration is shorter when full PEMT subtitles and HT subtitles are displayed than when RAW subtitles are displayed.

Table 8. Average Fixation Duration [ms] in Both AOIs of Each Group.

Table 8 shows that, regarding average fixation duration, Group RAW had the highest value in AOI_SUB (314.18), while Group PE had the lowest (258.94). Group HT had the highest value in AOI_IMA (398.61), while Group PE had the lowest (335.81). Therefore, this result only partly supports the hypothesis. Overall, three statistically significant differences for the average fixation duration measure were found: for AOI_SUB RAW vs PE, and for both AOI_SUB and AOI_IMA RAW vs HT.

4. Summary

Table 9 presents a summary for all results of ANOVA and means. The letter 'Y' indicates a statistically significant difference between the two groups, while 'N' means the opposite. The symbol '√' means the result supports the corresponding hypothesis, and the symbol 'x' means the result does not support the corresponding hypothesis.

Table 9. Summary for results of ANOVA and means.

According to the ANOVA results, it can be seen that the three group pairs return no statistically significant differences in most cases. In regard to Group RAW vs. Group

PE, the only statistically significant difference lies in the average fixation duration in AOI_SUB. The mean of average fixation duration in AOI_SUB for Group PE is 258.94, while that for Group RAW is 314.18. According to the means, most hypotheses are supported by the results, except the glance count in AOI_IMA and the fixation count in AOI_SUB. As mentioned in Section 3.3.1, there is a statistically significant difference between the comprehension testing score of the two groups, and Group PE outperformed Group RAW at both comprehension test and attitude survey. Relating all the results to the three 'R's in the reception model, we can see that Group PE performed better in 'Reaction' and 'Repercussion' than Group RAW. In regard to 'Response', Group PE performed partially better than Group RAW. Therefore, it can be concluded that, overall, participants who were offered full PEMT subtitles scored better on our reception metrics than those who were offered raw machine translated subtitles.

Group RAW and Group HT had statistically significant differences in half of the measures. Group HT outperformed Group RAW in comprehension testing, but their scores are not significantly different (see Section 3.3.1). The attitude survey result suggests that Group RAW had a better attitude towards the subtitles than Group HT (see Section 3.3.2), which goes against our hypothesis.

Group PE and Group HT return no statistically significant differences in most cases, the only exception is the fixation count in AOI_SUB. According to the means, Group PE outperformed Group HT in half of the measures. In addition, Group PE scored better than Group HT in comprehension testing, though the difference was not significant (see Section 3.3.1). Also, Group PE had a better attitude towards the subtitles than Group HT (see Section 3.3.2). Undoubtedly, this result confounds expectations.

It can be seen that Group HT, which was expected to have higher reception metrics than Group PE and Group RAW, returned some results that do not support

hypotheses. It has to be emphasized that all the hypotheses were built upon the premise that the quality of the subtitles would increase as we go from RAW to PEMT to HT. This premise was built on the intuition that human translation would outperform MT with PE, if the human translator is a good one. It has to be noted that the human translator is an English teacher rather than professional translator, and there is of course no guarantee that non-professional human translation can reach a quality level equal to that of a professional. For further analysis, a quality assessment for full PEMT subtitles and HT subtitles is imperative.

It has to be noted that the attitude survey of all groups reveals that most participants had a positive attitude towards their subtitles if we take 'Strongly agree' and 'Agree' as positive, and 'Strongly disagree' and 'Disagree' as negative (Group RAW: 75% positive, 6.49% negative; Group PE: 82.32% positive, 4.64% negative; Group HT: 66.66% positive, 9.18% negative).

5. Conclusion

This paper offers a contribution to the sparse research into machine translated subtitles for MOOCs. We test 20 hypotheses in total, 16 based on eye-tracking data and four on questionnaire data. Table 9 shows that among the 16 hypotheses, seven are not supported by the results (three in RAW vs. PE & HT, four in PE vs. HT). Half of the remaining four hypotheses were supported by the results, the other half, which involved Group HT, were not.

On the whole, results show that participants who were offered full PEMT subtitles scored better on our reception metrics than those who were offered raw MT subtitles, but not significantly. As mentioned before, the quality of the raw MT subtitles was relatively good. In regard to the participants who were offered HT subtitles, they

did not perform better than the other two groups, which is certainly not in support of our hypotheses and deserves further investigation. Notwithstanding this, most participants held a positive attitude towards the subtitles regardless of their type, which means MT can help in translating English subtitles for MOOCs into Chinese, and may contribute to the development of MOOCs in China in the long term.

This research hopefully provides empirical data for reception studies on machine-translated subtitles, albeit on a limited scale. We are fully aware that this research can be improved and expanded in many ways, for example, the number of participants would ideally be increased. Apart from that, some unexpected results of the experiment are worthy of further investigation, especially those regarding the HT subtitles. As we emphasized, our HT subtitles were produced by a non-professional translator. A quality assessment for HT and PEMT subtitles is underway.

Notes:

1. See: https://forum.opensubtitles.org/viewtopic.php?f=1&t=1969
2. Gambier's original model focuses on the core concepts rather than how they might be operationalized in an eye-tracking context.
3. The terminology adopted in this study is used by the eye-tracker manufacturer. Note, 'glances' are referred to as 'visits' in other eye-tracking studies.
   Fixation Count: number of fixations inside the AOI (SMI, 2015, p. 295).
   Glance Duration: Saccade duration for entering the object + sum of all fixation durations and saccade durations before the eyes begin to leave the AOI = dwell time + duration of saccade entering AOI (SMI, 2015, p. 280).
   Glance Count: number of glances to a target (saccades coming from outside) within a certain period (SMI, 2015, p. 295).
4. Average Fixation Duration: sum of durations of all fixations divided by number of fixations in the trial (SMI, 2015, p. 287).
5. Website of the English Proficiency Index: http://liuxue.ef.com.cn/epi/
6. See: https://www.letsmt.eu/Bleu.aspx
7. Website of Cambridge Assessment English: http://www.cambridgeenglish.org/test-your-english/general-english/
8. See: https://www.coe.int/en/web/common-european-framework-reference-languages/
9. 'Free student t-value Calculator', Free Statistics Calculators. https://www.danielsoper.com/statcalc/calculator.aspx?id=10 (accessed Jan 12, 2018)
10. See: https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/what-is-dark-and-bright-pupil-tracking/
11. This type of one-way ANOVA is carried out for all measures in the following sections. For reasons of space the analyses are not presented in detail, and only statistically significant differences are reported. A summary of results is provided in Table 9. Full details are available in the forthcoming thesis of one author.

References:

Alves, F., Gonçalves, J.L. & Szpak, K. (2012, December). Identifying instances of processing effort in translation through heat maps: An eye-tracking study using multiple input sources. In *COLING 2012. Proceedings of the 24th international conference on computational linguistics* (pp. 5-20). Mumbai, India: International Committee on Computational Linguistics.

Aranberri, N., Labaka, G., de Ilarraza, A.D. & Sarasola, K. (2014, October). Comparison of post-editing productivity between professional translators and lay users. In O'Brien, S., Simard, M., & Specia, L. (Eds.), *Proceedings of the third workshop on post-editing technology and practice* (pp. 20-33). Vancouver: Association for Machine Translation in the Americas.

Armstrong, S., Caffrey, C., & Flanagan, M. (2006, May). Translating DVD subtitles from English-German and English-Japanese using example-based machine translation. In Carroll, M., Gerzymisch-Arbogast, H., & Nauert, S. (Eds.), *Proceedings of the Marie Curie Euroconferences MuTra: audiovisual translation scenarios* (pp. 1-12). Copenhagen: EU High Level Scientific Conference Series.

Beaven, T., Comas-Quinn, A., Hauck, M., De los Arcos, B., & Lewis, T. (2013). The Open Translation MOOC: creating online communities to transcend linguistic barriers. *Journal of Interactive Media in Education*, *2013*(3), 1-14.

Blignaut, P., & Wium, D. (2014). Eye-tracking data quality as affected by ethnicity and experimental design. *Behavior Research Methods*, *46*(1), 67-80.

Caffrey, C. (2009). *Relevant abuse? Investigating the effects of an abusive subtitling procedure on the perception of TV anime using eye tracker and questionnaire* (Doctoral dissertation). Dublin City University, Dublin, Ireland. Retrieved from http://doras.dcu.ie/14835/

Castilho, S. (2016). *Measuring acceptability of machine translated enterprise content* (Doctoral dissertation). Dublin City University, Dublin, Ireland. Retrieved from http://doras.dcu.ie/21342/

Castilho, S., O'Brien, S., Alves, F., & O'Brien, M. (2014, June). Does post-editing increase usability? A study with Brazilian Portuguese as Target Language. In M. Tadić, P. Koehn, J. Roturier & A. Way (Eds.), *EAMT 2014. Proceedings of the seventeenth annual conference of the European Association for Machine Translation* (pp. 183-190). Dubrovnik, Croatia: European Association for Machine Translation.

China Education Centre Ltd. (2018, May 15). Project 211 and 985. Retrieved from https://www.chinaeducenter.com/en/cedu/ceduproject211.php

Doherty, S. (2012). *Investigating the effects of controlled language on the reading and comprehension of machine translated texts: a mixed-methods approach* (Doctoral dissertation). Dublin City University, Dublin, Ireland. Retrieved from http://doras.dcu.ie/16805/

Etchegoyhen, T., Bywood, L., Fishel, M., Georgakopoulou, P., Jiang, J., van Loenhout, G., del Pozo, A., Sepesy Maucec, M., Turner, A., & Volk, M. (2014, May). Machine Translation for Subtitling: A Large-Scale Evaluation. *LREC 2014. Proceedings of the Ninth*

International Conference on Language Resources and Evaluation (pp. 46-53). Reykjavik, Iceland: European Language Resources Association.

Filizzola, T. (2016). *Italians' perception and reception of British stand-up comedy humour with interlingual subtitles* (Doctoral dissertation). University College London, London, UK. Retrieved from http://discovery.ucl.ac.uk/1537589/

Flanagan, M. (2009, November). Using example-based machine translation to translate DVD Subtitles. In Forcada, M, & Way, A (Eds.), *Proceedings of the 3rd international workshop on example-based machine translation* (pp. 85-92). Dublin: Centre for Next Generation Localisation.

Gambier, Y. (2009). Perception and reception of audiovisual translation: Implications and challenges. In Che Omar, H., Haroon, H., & Abd Ghani, A. (Eds.), *The 12th international conference on translation: the sustainability of the translation field* (pp. 40-57). Kuala Lumpur, Malaysia: Malaysian Translators Association.

HCR 慧辰资讯. (2016, October 11). 深度：2016 年中国慕课行业研究白皮书. Retrieved from http://www.sohu.com/a/115837986_400678

ISO. (2017, May). ISO 18587:2017: Translation services – Post-editing of machine translation output – Requirements. Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:18587:ed-1:v1:en

Kordoni, V., van den Bosch, A. P. J., Kermanidis, K. L., Sosoni, V., Cholakov, K., Hendrickx, I. H. E., & Huck, M. (2016, May). Enhancing Access to Online Education: Quality Machine Translation of MOOC Content. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., & Piperidis, S. (Eds.), *LREC 2016. Proceedings of the 10th international conference on language resources and evaluation* (pp. 16-22). Portorož, Slovenia: European Language Resources Association.

Lison, P., & Tiedemann, J. (2016, May). OpenSubtitles2016: extracting large parallel corpora from movie and TV subtitles. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., & Piperidis, S. (Eds.), *LREC 2016. Proceedings of the 10th international conference on language resources and evaluation* (pp. 923-929). Portorož, Slovenia: European Language Resources Association.

MOE. (2018, January 19). Ministry of education launches 490 national selected online open courses. Retrieved from http://en.moe.gov.cn/News/Top_News/201801/t20180119_325124.html

MOOC 学院. (2014, August 11). 2014 年慕课学习者调查报告. Retrieved from https://mooc.guokr.com/post/610674/

Perego, E., Del Missier, F., Porta, M., & Mosconi, M. (2010). The cognitive effectiveness of subtitle processing. *Media Psychology*, *13*(3), 243-272.

Plitt, M., & Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics*, *93*, 7-16.

Schuster, M., Johnson, M. & Thorat, N. (2016, November 22). Zero-Shot Translation with Google's Multilingual Neural Machine Translation System. Retrieved from https://ai.googleblog.com/search/label/Google%20Translate

Secară, A. (2017). Cn U read ths? The reception of txt language in subtitling. In D. Kenny (Ed.), *Human Issues in Translation Technology* (pp. 167-188). London: Routledge.

SMI. (2015). BeGaze Manual Version 3.5. Teltow, Germany: SMI.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006, August). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th conference of the association for machine translation in the Americas* (pp. 223-231). Cambridge: Association for Machine Translation in the Americas.

Volk, M. (2008). The automatic translation of film subtitles. A machine translation success story? In Nivre, J., Dahllöf, M., & Megyesi, B. (Eds.), *Resourceful language technology: festschrift in honor of Anna Sågvall Hein* (pp. 202-214). Uppsala, Sweden: Uppsala University.

Wang, L., Zhang, X., Tu, Z., Way, A., & Liu, Q. (2016, May 22). Automatic construction of discourse corpora for dialogue translation. Retrieved from https://arxiv.org/abs/1605.06770

Appendix: Post-task questionnaire (administered in Chinese, translation into English by one author)

Part One: Comprehension questions

1. What counts as physical activity?
   A. Dancing
   B. Walking the dog
   C. Climbing the stairs
   D. Housework or gardening
   E. All of the above

2. Which of the following benefit of physical activity is not mentioned in the video?
   A. It can help reduce our risk of multiple diseases.
   B. It can help us to maintain healthy weight.
   C. It can help us to improve the quality of our life.
   D. None of the above.

3. Which of the following statement about physical activity is not right?
   A. It is any movement that uses energy.
   B. It is not structured.
   C. It is different from exercise.
   D. It is pursued for fitness benefits.

4. Which example below counts as a muscle strengthening activity?
   A. Running
   B. Yoga
   C. Lifting weights at the gym
   D. Gardening

5. Which of the following is not active transport?
   A. Walking
   B. Cycling
   C. Driving a car
   D. None of the above

6. Which of the following is not caused by sedentary behaviour?
   A. Obesity
   B. Type 2 diabetes
   C. Eye disease
   D. Some cancers
   E. Premature mortality

7. Which of the following is not exercise?
   A. Training to take part in a marathon
   B. Walking along the pavement
   C. Going to gym for strength training
   D. Going to a yoga class

8. According to physical activity guidelines, children should do 60 minutes of activity every day, that's _____
   A. Running
   B. Jumping
   C. Both of the above

9. According to physical activity guidelines, adults should _____
   A. have moderate activity at least 150 minutes every week
   B. have strength building activities
   C. minimize the time they spend sitting
   D. All of the above

10. True or false: Doing moderate activity is enough to meet the physical activity guidelines.
    A. True
    B. False

11. True or false: If you want to do physical activity only 10 minutes a day, 5 minutes a day, that's not going to be helpful.
    A. True
    B. False

12. True or false: Children get more active as they age.
    A. True
    B. False

13. True or False: Physical activity is a very specific subset of exercise.
    A. True
    B. False

Part Two: Attitude questions

14. The subtitles allow me to fully understand the contents of the MOOC.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

15. The subtitles are useful to me.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

16. The subtitles are easy to understand.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

17. Interacting with the subtitles does not require a lot of my mental effort.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

18. I would find it easy to get the information I need from subtitles.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

19. The subtitles are clear and understandable.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

20. I enjoyed reading the subtitles.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

21. I am satisfied with the subtitles.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

22. If I have a chance, I would use machine translation to translate English subtitles in the future, because I know it will do a good job.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

23. I would recommend machine translation to my friends if they need to translate subtitles.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

I could comprehend the subtitles…

24. If there was no one around to tell me what to do as I go.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

25. If I could call someone for help if I got stuck.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

26. If I had a lot of time.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

27. If I had just the built-in help facility for assistance.
☐ Strongly agree ☐ Agree ☐ Neutral ☐ Disagree ☐ Strongly disagree

Table 1. Reception Model and Associated Measurement Tools (adapted from Gambier 2009)

Element	Related to	Reflected in	Measured by
Response	Perceptual decoding	Attentional processes	Eye-tracking
Reaction	Psycho-cognitive level	Processing effort and comprehension	Eye-tracking and comprehension testing
Repercussion	Attitudinal issues and sociocultural dimensions	Attitudes and beliefs	Attitude questions

Table 2. Comprehension testing score of the participants reporting Level C competence.

Participant	English score (full score: 25)	Comprehension testing score (full score: 13)
P26 (RAW)	21	8
P34 (RAW)	22	9
P36 (RAW)	20	8
P38 (RAW)	20	10
P49 (HT)	21	7
P51 (HT)	20	12
P54 (HT)	20	9

Table 3. Comprehension testing score per group.

	Max	Min	Mode	Mean	SD
Group PE (24)	13	6	10, 11	9.58	1.74
Group RAW (22)	11	6	8	8.55	1.37
Group HT (15)	12	6	9, 10, 11	9.47	1.85

Table 4. Attitude survey results per group.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Group PE (24)	33.04% (114)	49.28% (170)	13.04% (45)	4.35% (15)	0.29% (1)
Group RAW (22)	17.21% (53)	57.79% (178)	18.51% (57)	6.49% (20)	0
Group HT (15)	27.05% (56)	39.61% (82)	24.16% (50)	9.18% (19)	0

Note: One participant in Group HT failed to answer Statements 25, 26, 27, and 28. Hence, the number of respondents for the four questions in this group is 14.

Table 5. Glance count in both AOIs of each group.

Group PE			Group RAW			Group HT		
	AOI_SUB	AOI_IMA		AOI_SUB	AOI_IMA		AOI_SUB	AOI_IMA
P01	150	140	P25	163	169	P47	119	118
P02	163	146	P26	151	143	P48	119	119
P04	107	109	P28	175	175	P51	197	163
P06	144	146	P29	_	_	P52	_	_
P09	121	124	P30	119	89	P53	152	142
P10	178	184	P33	115	115	P54	102	168
P11	132	132	P35	128	126	P55	117	114
P12	135	135	P36	140	142	P56	138	139
P14	_	_	P37	186	187	P57	106	83
P17	186	168	P38	132	131	P58	163	168
P23	171	173	P39	173	162	P59	139	137
P24	79	81	P40	129	127	P60	112	104
			P43	202	157	P61	145	146
			P46	105	103			
Mean	142.36	139.82		147.54	140.46		134.08	133.42
SD	32.19	29.44		30.01	29.01		27.51	26.58

Table 6. Fixation count for both AOIs for each group.

Group PE			Group RAW			Group HT		
	AOI_SUB	AOI_IMA		AOI_SUB	AOI_IMA		AOI_SUB	AOI_IMA
P01	758	492	P25	670	501	P47	679	367
P02	959	370	P26	636	431	P48	334	509
P04	800	270	P28	666	416	P51	687	536
P06	464	573	P29	_	_	P52	_	_
P09	390	468	P30	809	206	P53	372	458
P10	656	637	P33	873	275	P54	176	530
P11	1051	331	P35	712	289	P55	930	292
P12	885	300	P36	660	450	P56	810	411
P14	_	_	P37	617	495	P57	520	174
P17	733	450	P38	513	287	P58	454	333
P23	829	484	P39	712	383	P59	353	692
P24	764	186	P40	722	297	P60	531	246
			P43	601	331	P61	359	413
			P46	797	208			
Mean	753.55	414.64		691.38	351.46		517.08	413.42
SD	195.42	135.92		95.93	101.32		221.23	143.36

Table 7. Glance Duration [s] in both AOIs of each group.

	Group PE			Group RAW			Group HT	
	AOI_SUB	AOI_IMA		AOI_SUB	AOI_IMA		AOI_SUB	AOI_IMA
P01	207.20	174.27	P25	194.36	139.90	P47	240.44	156.36
P02	264.03	120.87	P26	189.18	141.69	P48	120.34	276.36
P04	274.53	124.24	P28	183.58	162.40	P51	211.62	174.85
P06	131.73	236.04	P29	_	_	P52	_	_
P09	109.79	199.47	P30	287.82	79.25	P53	101.88	148.89
P10	174.45	230.13	P33	296.80	85.02	P54	44.53	149.88
P11	318.21	97.08	P35	264.83	115.54	P55	311.00	92.68
P12	271.81	126.92	P36	226.88	163.95	P56	210.14	138.31
P14	_	_	P37	186.88	205.32	P57	203.89	73.74
P17	172.17	108.34	P38	247.07	165.10	P58	175.12	192.74
P23	227.81	172.67	P39	226.06	148.74	P59	85.87	253.84
P24	279.04	91.61	P40	230.58	120.87	P60	235.81	167.04
			P43	193.66	131.12	P61	120.55	268.46
			P46	311.13	70.07			
Mean	220.98	152.88		233.76	133.00		171.77	174.43
SD	67.26	52.22		44.73	38.80		77.47	64.41

Table 8. Average Fixation Duration [ms] in Both AOIs of Each Group.

Group PE			Group RAW			Group HT		
	AOI_SUB	AOI_IMA		AOI_SUB	AOI_IMA		AOI_SUB	AOI_IMA
P01	242	319.7	P25	264.2	227.6	P47	319.9	390.1
P02	250.4	285.9	P26	270.2	285.1	P48	333.7	517.9
P04	318.7	426	P28	250.7	359	P51	202.7	239.9
P06	250.3	376.3	P29	_	_	P52	_	_
P09	235.6	388.7	P30	320.5	336.4	P53	236.7	277.1
P10	230.4	311.7	P33	321.2	275.3	P54	209.8	228.1
P11	272	248.6	P35	347.5	364.9	P55	304.9	276.2
P12	283	383.7	P36	315.9	331.4	P56	234.8	300.2
P14	_	_	P37	273.3	378.4	P57	371.7	401.1
P17	177.2	180.9	P38	461	543.9	P58	362.6	551.7
P23	244.5	321.5	P39	292.3	358.7	P59	215.7	333
P24	344.5	450.9	P40	297.4	374.3	P60	422.5	646
			P43	302.9	368.2	P61	310.1	622
			P46	367.2	306.1			
Mean	258.94	335.81		314.18	346.87		293.76	398.61
SD	45.03	79.53		55.22	74.61		72.75	149.80

Table 9. Summary for results of ANOVA and means.

		AOI	Glances Count	Fixation Count	Glance Duration	Average Fixation Duration
ANOVA	RAW vs. PE	SUB	N	N	N	Y
		IMA	N	N	N	N
	RAW vs. HT	SUB	N	Y	Y	Y
		IMA	N	N	N	Y
	PE vs. HT	SUB	N	Y	N	N
		IMA	N	N	N	N
Means	RAW vs. PE & HT	SUB	> √	> x	> √	> √
		IMA	< x	< √	< √	> x
	PE vs. HT	SUB	> √	> √	> √	> x
		IMA	< x	< x	< √	< x