Reading Comprehension of Machine Translation Output: What Makes for a Better Read?

Sheila Castilho
ADAPT Centre
Dublin City University
sheila.castilho@adaptcentre.ie

Ana Guerberof Arenas
ADAPT Centre/SALIS
Dublin City University
ana.guerberof@adaptcentre.ie

Abstract

This paper reports on a pilot experiment that compares two different machine translation (MT) paradigms in reading comprehension tests. To explore a suitable methodology, we set up a pilot experiment with a group of six users (with English, Spanish and Simplified Chinese languages) using an English Language Testing System (IELTS), and an eye-tracker. The users were asked to read three texts in their native language: either the original English text (for the English speakers) or the machine-translated text (for the Spanish and Simplified Chinese speakers). The original texts were machine-translated via two MT systems: neural (NMT) and statistical (SMT). The users were also asked to rank satisfaction statements on a 3-point scale after reading each text and answering the respective comprehension questions. After all tasks were completed, a post-task retrospective interview took place to gather qualitative data. The findings suggest that the users from the target languages completed more tasks in less time with a higher level of satisfaction when using translations from the NMT system.

1 Introduction

Recently, there has been an increase in Neural Machine Translation (NMT) research as contemporary hardware supports much more powerful computation during the creation process. Research on the translation quality of NMT engines show that, in general, when compared against Statistical Machine Translation (SMT) engines, the output quality of NMT systems is higher when measured using automatic metrics (Bahdanau et al., 2014; Jean et al., 2015; Bojar et al., 2016; Koehn and Knowles, 2017). However, results are not as positive when human evaluators compare these outputs (Bentivogli et al., 2016; Castilho et al., 2017a; Castilho et al., 2017b).

Human evaluation of MT output, although not always implemented in quality evaluation, has been increasingly endorsed by researchers who acknowledge the need for human assessments. Some of the most commonly-used manual metrics are fluency and adequacy, error analysis, translation ranking, as well as post-editing effort. Despite the considerable focus on MT quality evaluation, the impact of MT on the end user has been under-researched. Measuring the usability of MT output allows for identification of the impact that the translation might have on the end user (Castilho et al., 2014). With the intention of exploring the cognitive effort required to read texts originating from SMT and NMT engines by the end users of those texts, we set-up a pilot experiment that aims to measure the reading comprehension of Spanish and Simplified Chinese users of texts produced by both paradigms using an eye-tracker (using the English users' data as a baseline).

The remainder of this paper is organized as follows: in Section 2, we survey the existing literature concerning reading comprehension for MT evaluation and the use of eye-tracking techniques for translation assessment; in Section 3, we describe the research questions and hypotheses which guide this pilot experiment, as well as the methodology applied to carry out the experiment with English

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 79–88
Alacant, Spain, May 2018.

(EN), Spanish (ES) and Simplified Chinese (ZH) native speakers; the results are discussed in Section 4, and finally, in Section 5, we draw the main conclusions of the pilot study and outline promising avenues for future work.

2 Related Work

2.1 Reading Comprehension for Machine Translation Evaluation

Despite the considerable focus on MT quality evaluation, there has not been much research focused on the impact of MT on the end user. With the current shift of paradigm in the MT landscape, it has become essential to also test the reading comprehension of NMT models by the end users of those translations. A few studies have attempted to measure reading comprehension (Scarton and Specia, 2016) and usability of MT output. Tomita et al. (1993) use reading comprehension tests to compare different MT systems. The content for reading and comprehension was extracted from an English proficiency exam and then translated into Japanese via three commercial MT systems as well as through the process of human translation. Sixty native speakers of Japanese were asked to read the text and answer the questions. The authors show that reading comprehension is a valid evaluation methodology for MT; however, their experiment only takes into consideration the informativeness, i.e. the number of correct answers for the comprehension questions.

Fuji (1999) proposes reading comprehension tasks in order to measure informativeness and, moreover, the author adds comprehensiveness and fluency to the evaluation measures. The content used comprises several texts from official examinations of English language designed for Japanese students. Participants were asked to read the text, answer the comprehension questions and judge how comprehensible and how fluent the text is, using a 4 point scale. Following on from this, Fuji et al. (2001) examined the "usefulness" of machine-translated text from two commercial MT systems compared to the English version. The experiment consisted of participants reading the texts and answering comprehension questions. The authors claim that presenting the source with the MT output results in higher comprehension performance.

Jones et al. (2005) ask 84 English native speakers to answer questions from a machine-translated and human-translated version of the Defense Language Proficiency Test for Arabic language. Task time and subjective rating were also measured. Their results suggest that MT may enable a limited working proficiency but it is not suitable for a general professional proficiency.

Usefulness, comprehensibility, and acceptability of MT technical documents are examined by Roturier (2006). The author claims that a text is deemed useful when readers are able to solve their problem with the help of the translation. The study uses a customer satisfaction questionnaire to determine whether controlled English rules can have a significant impact from a Web users perspective. The main drawback of Roturiers approach is that there is no task being performed by the end user as the methodology consists of an online questionnaire.

2.2 Eye tracking in Translation Research

Doherty and O'Brien (2012) is the first study to use eye-tracking techniques to measure the usability of translated texts via the end user. They conduct a study to compare the usability of raw machine-translated output for four target languages (Spanish, French, German and Japanese) against the usability of the source content (English). The result of this first phase compared the machine-translated group against the source group, and found significant difference for goal completion, efficiency, and user satisfaction between the source and the MT output. In the second phase of the study, Doherty and O'Brien (2014) analyse the results according to target languages compared to the source. The results show that the raw MT output scores lower for usability measurements, requiring more cognitive effort for all target languages when compared with the source language content.

Stymne et al. (2012) present a preliminary study using eye tracking as a complement to MT error analysis. In this methodology, although the main focus is to identify and classify MT errors, a comprehension task is also applied. For the perception questions, the human translation scored better than all the MT options. For both perceived and actual reading comprehension questions, their results show that participants are more efficient when using the MT output of a system trained using a large corpus. Regarding gaze data, MT errors are associated with both longer gaze times and more fixations than correct passages, and average gaze time is dependent on the type of errors which may sug-

gest that some error types are more disturbing for readers than others.

Klerke et al. (2015) present an experimental eye-tracking usability test with text simplification and machine translation (for both the original and simplified versions) of logic puzzles. Twenty native speakers of Danish were asked to solve and judge 80 different logic puzzles while having their eye movements recorded. A greater number of fixations on the MT version of the original text (with no simplification) was observed and participants were less efficient when using the MT version of the original puzzles; however, the simplified MT version seemed to ease task performance when compared to the original English version.

Castilho et al. (2014) had two groups of 9 users each performing tasks using either the raw MT or the post-edited version of instructions for a PC-based security product, and cognitive and temporal effort indicators were gathered using an eye-tracker. Their results show that lightly post-edited instructions present a higher level of usability when compared to raw MT. Building on this, Castilho and O'Brien (2016) perform similar experiments with German and English native speakers, with instructions for spreadsheet software. Results show that the post-editing group is faster, more efficient, and more satisfied than the MT group. No significant differences appear in cognitive effort between raw and post-edited instructions, but differences exist between the post-edited versions and the source language. Moreover, the authors claim that the cognitive data should not be viewed in isolation, and highlight the importance of collecting qualitative data for measuring usability. Finally, Castilho (2016) extended previous experiments using Simplified Chinese, Japanese, German and English for the same set of instruction of the spreadsheet software. Results show that participants who used the post-editing instructions were more effective, more efficient, and faster than participants who used the raw MT instructions, especially for Simplified Chinese and German. Another interesting finding is that the source mostly did not differ from the post-editing groups, suggesting that the post-editing output is of equivalent quality. Regarding satisfaction, the author reports that German participants who use the MT instructions, even though they are able to successfully perform more tasks than other MT groups, are the least satisfied with the instructions, while the Japanese participants do not present any difference between the MT and post-editing groups for satisfaction even though the MT group was the least efficient. The author notes that these findings are likely to be related to cultural characteristics, as the Japanese participants are more tolerant and less likely to complain. Another interesting finding is that all groups, including the English-speaking participants, suggest that the instructions need improvements.

Finally, Jordan-Nez et al. (2017) compare three MT systems for assimilation, namely Systran (hybrid corpus based and rule-based MT); Google Translate (at the time of the experiment, a SMT system); and Apertium (a rule-based system), against professional translations. Results show that the MT output into a language in the same family as the readers first language may facilitate comprehension of texts originally written in a language from a different family. The authors note, however, that the level of usefulness depends on the field and on the MT system used as well as on the level of speciality.

Following previous work, we expect that the MT system that shows closer efficiency measures to the source text and lower task time, as well as lower cognitive effort indicators, is more likely to be rated higher for the satisfaction.

3 Methodology

Hypothesis and Research Questions As mentioned in Section 1, the primary aim of this experiment is to gather more information about the user experience when reading for comprehension machine-translated texts. With this aim in mind, we identified the following research questions:

RQ1: Which MT engine offers better efficiency to participants, i.e. with which one are they able to successfully answer more comprehension questions? Or with which one are they able to complete the tasks faster?

RQ2: To what extent are there differences in participants cognitive processes due to different engines (NMT and SMT)?

RQ3: What is the participants level of satisfaction with SMT and NMT when reading for comprehension?

Content and Design In order to answer the research questions, we measured participants reading comprehension according to the number of

correct answers (goal completion) to a set of comprehension questions about each text, and task time. Eye-tracking fixation count and duration are also computed, as well as satisfaction indexes after each reading task. After all tasks were completed, we interviewed the participants by means of a semi-structured retrospective interview to gauge the understanding of the texts from a qualitative perspective.

For this pilot, we recruited two native speakers per language, a total of six participants (English, Spanish and Simplified Chinese languages). In this case, we used a sample of convenience. The participants were part of the student and staff body of Dublin City University. There were three female and three male participants, average age was 30.6 years, and all of them had received education to a post-graduate level. Half of them had previous experience in reading comprehension tests, either as part of their education or work. The Spanish and Simplified Chinese participants had a university level standard of English as they have taken English Proficiency tests and have been working and studying in an English-speaking country for some time.

As for the reading texts, two were taken from the International English Language Testing System (IELTS)[1] that measures English language proficiency by assessing four language skills: listening, reading, writing and speaking. IELTS has two types of tests: General and Academic. Since we were trying to assess the reception of raw output for a general user, we decided to use the General Training IELTS, reading modality, which contains a text and comprehension questions about that same text. The total number of words in the source content amounted to 1090 words.

The two English texts selected and their accompanying comprehension questions were then translated using Microsoft Translator Try and Compare feature[2] that allowed one to generate output in both SMT and NMT, and compare their quality. The first text (Text 2), entitled "Beneficial work practices for the keyboard operator", contained seven comprehension questions in which the users were required to choose the correct heading for each section of the text from a list of headings. The second text (Text 3), entitled "Workplace dismissals", contained five comprehension questions for which the users were required to match each description from a list with a correct term displayed in a box. One short text was also extracted from the IELTS website to be used as baseline. This baseline text (Text 1) was available in English, Spanish and Simplified Chinese on the IELTS website.[3] Moreover, ten questions in the style of the test (write True, False or Not given) were created in English for this baseline text and translated into Spanish by a Spanish translator and into Simplified Chinese by a native speaker. The baseline was used to test participants attention and reading comprehension with a human-translated version. The total number of words in the source baseline text amounted to 229 words. The baseline text was presented first followed by the Text 2 and Text 3 (SMT and NMT) which were randomised.[4] Figure 1 shows the set up of the task.

After each task (text and comprehension questions), four statements were presented (in English) in a three-point Likert scale (1- disagree, 2- neither agree or disagree, 3- agree) for the participants:

1. The subject of the text was easy to understand.
2. The language was easy to understand.
3. The question was easy to understand.
4. I was able to answer the question confidently.

The eye tracker used was a Tobii T60XL with the filter set for I-VT (Velocity-Threshold Identification), as this is the filter recommended by Tobii for reading experiments. The participants were recorded during the post-task interview using the Flashback application that allows recording of all movements, sounds, and webcam output on the computer. This retrospective post-task interview was designed so that participants could watch their recordings and give their feedback regarding the subject matter, language used, questions, and personal experience when completing the whole task.

[1] https://www.ielts.org

[2] The feature on the website has changed to a comparison between Microsoft's production and research engines. See https://translator.microsoft.com/neural.

[3] As this text was available on the target languages on the IELTS official website, we assume that the translations were either direct human translation of the source or they were comparable texts, i.e. texts with the same information but originally written in the target language.

[4] The same order of texts were presented for the English participants (Text 1, Text 2 and Text 3) but in the source EN language.

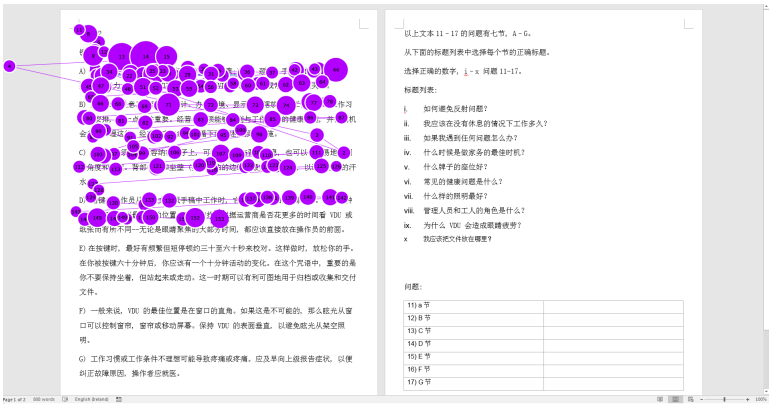

Figure 1: Task set-up

Texts	EN P02	EN P04	ES P01	ES P05	ZH P09	ZH P08	Av. per system/text Baseline/ source	SMT	NMT
Baseline	90	90	100	100	90	80	92	—	—
Text 2	57	100	85	71	85	71	79	71	85
Text 3	60	100	60	100	100	100	80	80	100

AV. per system/lg						
SOURCE	79		—	—	—	—
SMT	—	—	66		86	
NMT	—	—	93		93	

Figure 2: Goal Completion (%)

Texts	EN P02	EN P04	ES P01	ES P05	ZH P09	ZH P08	Av. per system/text Baseline/ source	SMT	NMT
Baseline	68	148	88	140	250	206	150	—	—
Text 2	346	657	417	552	502	360	502	456	459
Text 3	241	212	272	333	528	272	226	400	302

AV. per system/lg						
SOURCE	364		—	—	—	—
SMT	—	—	412		444	
NMT	—	—	375		387	

Figure 3: Task Time (in seconds)

4 Results

4.1 Comprehension

As mentioned previously, the baseline (Text 1) contained 10 questions, while Text 2 contained 7 questions, and Text 3 contained 5 questions. Goal completion is the number of successfully completed tasks, while task time is the total task time the participants needed to complete the tasks.

Goal Completion Figure 2 shows the results for goal completion for all participants (P01, P02, P04 and so on), where light gray cells are SMT while dark gray cells are NMT results. We can see that on average, participants who read the NMT text had a higher rate of goal completion (ES and ZH: 93%) when compared to the participants who read the SMT texts (ES: 66%, ZH: 86%), even when compared to participants who used the English source (79%). Interestingly, Simplified Chinese participants who used the SMT tests also had higher rates of goal completion when compared to the average for the English text.

When looking at the average score per system for each text (last column), participants of all languages had higher goal completion when reading Text 3 when compared to Text 2, which may indicate that Text 3 was easier to understand[5]. This is mentioned during the retrospective interviews by the participants (see Section 4.4).

Task Time Regarding the amount of time required for participants to read the texts and answer the comprehension questions, Figure 3 shows that,

[5]Text 3 contained 5 questions, whereas Text 2 contained seven question which could also have impacted goal completion

on average, participants who read the NMT output (ES: 375, ZH: 387) were faster than participants who read the SMT output (ES: 412, ZH: 444). Additionally, participants who used the NMT texts, for both ES and ZH, have closer average task time to participants who used the source text. Interestingly, the Simplified Chinese participants seemed to spend slightly more time on the task than the ES and EN participants, which could be related to the fact that the ZH participants were able to answer more questions correctly.

4.2 Eye-Tracking Data

As previously mentioned, we used an eye tracker to collect empirical data to analyse cognitive effort. Due to the low number of participants for the first part of this study, it is not possible to report any statistically significant results. However, we believe that these preliminary results may indicate a tendency in cognitive effort between NMT and SMT.

Fixation Duration (FD) is the length of fixations (in seconds) within an area of interest (AOI). The longer the fixations are, the higher the cognitive effort may be expected. Figure 4 shows the results for the length of fixations. The average fixation duration per system indicates that SMT presents longer fixations (sum) when compared to the NMT system for both ES and ZH. However, the mean length does not seem to differ much, and, in fact, for ZH it presents a slightly shorter mean (0.25 secs) than the NMT system (0.26 secs). In general, ZH participants present longer FD mean results when compared to ES and EN for both systems, including for the baseline (Text 1), which correlates with the time ZH participants spent on tasks (Figure 3).

Fixation Count (FC) is the total number of fixations within an AOI. The more there are, the higher the cognitive effort is deemed to be. The average FC per system for each language in Figure 5 indicates that, in general, SMT presents a higher number of fixations when compared to the fixation for the NMT system for both ES and ZH languages. Interestingly, ZH does not show higher means for FC as previously observed for FD. In fact, ZH participants show lower FC when compared to Spanish, and in the case of NMT, lower than the English as well.

4.3 Satisfaction

As stated previously in Section 3, after the participants had completed each text and answered the comprehension questions, they were presented with four statements that measured their level of satisfactions with the subject of the text (the subject of the text was easy to understand), language (the language was easy to understand), questions (the question was easy to understand) as well as their perceived confidence (I was able to answer the question confidently) when answering the questions, in a 3-point Likert scale (3-agree, 1-disagree). Figure 6 presents the results for all languages.

In Figure 6, the average per system for each language shows that participants who used the EN texts have the highest satisfaction levels (2.56). For ES, participants who used the NMT system seem to be slightly more satisfied (1.6) than participants who used the SMT system (1.5). The same pattern can be seen in the ZH participants' satisfaction scores, the average for the NMT was considerably higher (2.37) than for the SMT system (1.37). This is in line with the task time (Figure 1) and goal completion (Figure 2) for the ZH language, in which participants were able to complete 93% of the tasks in an average of 387 secs using NMT translations, while using SMT translation they were able to complete 86% of the tasks in over 444 seconds. These results also illustrate the comments from the participants presented in the following section.

4.4 Retrospective Interviews

To triangulate the data from the eye-tracker and the statements presented to the participants after each task is completed (satisfaction scores), and obtain a more accurate account of the differences between SMT and NMT in reading comprehension tests, we carried out retrospective interviews with all participants. After each participant had completed the three tasks, we replayed the video of their eye movements in the Replay window of Tobii Studio, and recorded these interviews using Flashback as part of a Retrospective Think Aloud protocol. We asked the participants to watch the video showing their fixations on the screen and to describe freely their recollection of what they were thinking or doing at that time in the exercise. We clarified that they should not be worried about any grammar mistakes since four out of six of the

Texts	EN				ES				ZH				Av. per system/text		
	P02		P04		P01		P05		P09		P08				
	MEAN	SUM	MEAN	SUM	MEAN	SUM	MEAN	SUM	MEAN	SUM	MEAN	SUM	Baseline / source	SMT	NMT
Baseline	0.16	43.52	0.26	87.83	0.21	61.33	0.18	83.67	0.25	295.96	0.26	421.46	0.22	—	—
Text 2	0.18	252.63	0.23	525.98	0.18	332.01	0.17	387.32	0.26	219.11	0.26	421.46	0.21	0.22	0.22
Text 3	0.18	180.47	0.23	168.57	0.18	177.06	0.17	205.36	0.25	295.96	0.25	402.93	0.21	0.22	0.21

Av. per system/lg		EN		ES		ZH	
		MEAN	SUM	MEAN	SUM	MEAN	SUM
	SOURCE	0.20	1127.65**	—	—	—	—
	SMT	—	—	0.18	564.38	0.25	717.41
	NMT	—	—	0.18	537.37	0.26	622.04

Figure 4: Fixation Duration - in seconds.(** is the sum for both EN participants for both Text 2 and 3)

Texts	EN		ES		ZH		Av. per system/text		
	P02	P04	P01	P05	P09	P08			
	SUM	SUM	SUM	SUM	SUM	SUM	Baseline / source	SMT	NMT
Baseline	269	334	298	466	612	712	449	—	—
Text 2	1371	2323	1810	2217	851	1636	1847	1927	1331
Text 3	991	738	978	1211	1200	1588	865	1089	1400

Av. per system/lg		EN		ES		ZH	
		MEAN	SUM	MEAN	SUM	MEAN	SUM
	SOURCE	1355.7	5423.0**	—	—	—	—
	SMT	—	—	1597.5	3195.0	1418.0	2836.0
	NMT	—	—	1510.5	3021.0	1219.5	2439.0

Figure 5: Fixation Count (** is the sum for both EN participants for both Text 2 and 3)

Texts	Statements	EN		ES		ZH		Av. per system/text		
		P02	P04	P01	P05	P09	P08	Baseline/ source	SMT	NMT
Baseline	Subject	3	3	3	3	3	3	3.0	_	_
	Language	3	3	3	3	3	3	3.0	_	_
	Questions	3	3	3	3	3	2	2.8	_	_
	Confidence	2	3	3	3	3	3	2.8	_	_
Text 2	Subject	2	3	2	3	1	1	2.5	2.0	1.5
	Language	1	3	1	1	3	1	2.0	1.0	2.0
	Questions	3	1	3	1	3	3	2.0	2.0	3.0
	Confidence	2	3	2	1	1	1	2.5	1.0	1.5
Text 3	Subject	2	3	1	1	1	2	2.5	1	1.5
	Language	3	3	2	1	2	3	3.0	2	2
	Questions	3	3	1	1	1	3	3.0	1.0	2.0
	Confidence	3	3	2	2	1	3	3.0	1.5	2.5

AV. per system/lg							
	SOURCE	2.56		_	_	_	_
	SMT	_	_	1.5		1.4	
	NMT	_	_	1.6		2.4	

Figure 6: Ratings of Satisfaction (the higher score, the better)

participants did not have English as their mother tongue, the language in which the interviews were conducted. At the time of writing this paper, we have not completed a full qualitative analysis of these interviews, that is transcription and coding of the recordings, therefore what we provide here is a summary of the preliminary results.

All participants in all languages indicated that

Text 1 (the baseline text: original English or human translation) was easy to understand. They found the text to be short, the content easy to understand, and the language clear. Regarding Text 2, although most participants mentioned that it was more time consuming mainly due to the number of questions and options available (seven questions and ten options to choose from), their assessment of the language quality varied depending on the language and the type of engine used for this experiment. The same applies for Text 3, although the participants indicated that it was faster to complete because there were fewer questions and they already knew the dynamic of the exercises.

In the case of the English-speaking participants, they did not mention any aspects of the language or content that they found particularly difficult, although one participant (P02) had difficulties with the coding system to answer the questions in Text 1 (True, False, Not given). This participant also mentioned that he was not happy with certain commas or double negatives on Text 2. He did not find any linguistic issues on Text 3. The other English participant, P04, found the language to be satisfactory.

If we look at the Spanish language, P01 mentioned that Text 2 (NMT engine) was "more confusing" than Text 1 (Human translation). There were keywords that were "tricky" and she thought they were probably wrong, such as *sostenedor* instead of *atril* for *holder*, also she mentioned words that seemed to be completely out of context, such as *hechizo* for spell. Regarding Text 3 (SMT), the participant said that it was "really, really tricky" and "the language was really difficult" not because of words but because of incorrect grammar, and she stated that sentences were difficult to understand. She commented that "there were times where it came to my mind that these were direct translations from English". Because of the incorrect translations provided by the engine (two English options were translated in the same way in Spanish by the SMT engine), the participant answered two questions incorrectly. Participant 5 mentioned that in Text 2 (SMT, in this case), he noticed grammar mistakes "straight away", and then he realised that "it was translated by a machine" as "almost every sentence had something wrong". He mentioned that, although he had to read the sentences several times to try and make sense of the meaning, the content was not difficult for him. On the other hand, he found Text 3 (NMT, in this case) easier because there were fewer questions to answer, but he also mentioned that Text 3 was machine-translated. He noticed a few grammar errors and inconsistencies. For example, he noticed *Despido sumario* and *Resumen despido* as a translation for *Summary Dismissal*, and *Constructivo Despido* and *Constructivo despido* for *Constructive dismissal*, and this created confusion when he was answering the comprehension questions. He thought that the language was more technical than in the other documents but at the same time that the questions were easier to answer. When asked if he saw any difference between Text 2 and Text 3, he said that he had no reasons to assume a different MT system was used.

Regarding the Simplified Chinese language, P08 stated that Text 2 (SMT in this case) was the most difficult text of the three. According to him, Text 2 "was not fluent", some words were "weird", and he had to guess a lot of the text by the context and the questions. For him, the first two paragraphs, for example, were difficult to understand. Therefore, both contents and language were difficult. Regarding Text 3 (NMT), P08 found that it was "in the middle of the three". The paragraphs were "better" and the questions were "clear". Although, the content was new to the participant, he found the language easier to understand in Text 3 than in Text 2 but worse than in Text 1, as "the words were correct", but the order was wrong, and there were also characters missing. As for P09, she found that the structure of Text 2 (NMT, in this case) was "okay" but she was not familiar with the topic. She thought the language was also "okay"; although there were errors and sometimes the vocabulary was incorrect, she could understand it. In this text, she found the headings difficult to place in the corresponding section. P09 found that Text 3 (SMT in this case) was the most difficult one. She understood that the text was about dismissals, but she found the language "strange", "totally unclear", "the structure was not that good" and it was "hard to understand". She found that Text 2 and Text 3 were stressful, especially Text 3. She commented that she could understand 60 percent of Text 2, but only 20 percent of Text 3.

In summary, the EN participants found Text 2 more cumbersome to resolve than Text 1 and Text 3, and therefore more time was required, but only P02 mentioned that the language was an issue and

that it could be improved in Text 2 with regards to commas and double negatives. This is very interesting as it suggests that the difficulties EN participants found in the source could have been translated in the target languages. For ES and ZH, the four participants found Text 1 (human translation) easy in content and language, while they were divided on Text 2 and Text 3. In Simplified Chinese, the texts translated with NMT, regardless of whether they were Text 2 or 3, were viewed as better linguistically than their counterparts translated with SMT, even when the NMT texts had certain terms or grammar turns that were wrong, and this influenced the participants' responses. In Spanish, one of the participants found the NMT option better linguistically, while the other participant found that both options were comparable and possibly came from the same MT system.

5 Conclusions and Future Work

The aim of this pilot experiment was to verify the methodology to measure the impact of the quality of two MT paradigms - NMT and SMT - on the end user. For that, we established three research questions regarding efficiency (goal completion), cognitive effort, and satisfaction.

Regarding RQ1 (Which MT engine offers better efficiency to participants?), results show that participants (Figure 2) in the two target languages - Spanish and Simplified Chinese - were able to complete more tasks successfully when using the NMT translated texts when compared to the SMT translations, as well as when compared to participants who used the original EN texts. Regarding the time spent to complete the texts, again, we noted that when using the NMT translations, participants were faster than when using SMT translations and, moreover, have task completion times closer to participants who used the English text than the results for SMT.

Regarding RQ2 (To what extent are there differences in participants cognitive processes due to different engines?), results for the FD (Figure 4) and FC (Figure 5) show that cognitive effort does not seem to differ much for ES, and presents a bit of mixed results for ZH, were FD are slightly longer for the NMT system, whereas FC are lower. We believe that with a greater number of participants, a clearer tendency would be observed.

Regarding our last research question (RQ3: What is the participants level of satisfaction with SMT and NMT when reading for comprehension?), participants rated NMT higher and also commented that the language in NMT texts was easier to understand in the post-task retrospective interviews. It is also necessary to point out that ES and ZH participants commented on the fact that the language in the human translation (Text 1) was easy to understand, while they struggled in certain sections in both NMT and SMT texts (Texts 2 and 3). This was not the case with EN participants that only made slight remarks on the quality of the English, but they did not mention any misunderstandings of the texts.

We are aware of the limitations of the results presented here since the number of participants was very low, and there were few texts for each MT system. Our next steps are to add more languages, especially those languages which have been showing greater improvement with NMT over the SMT paradigm, as well as gathering more participants. Another consideration to bear in mind is the nature of the texts; we noted that the combination of difficult text with easy questions and vice-versa could cloud the findings.

Furthermore, we believe that this research could benefit from computing more eye-tracking measures, such as visit count, which is the number of visits to an area of interest, as the shifts of attention between the questions and the text may be an indicator of cognitive effort (Castilho et al., 2014).

Acknowledgements: We would like to thank Dag Schmidtke from Microsoft Ireland and Joss Moorkens for invaluable help, and the participants for their support on this pilot experiment. This research was supported by the Edge Research Fellowship programme that has received funding from the European Unions Horizon 2020 and innovation programme under the Marie Sklodowska-Curie grant agreement No. 713567, and by the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded under the European Regional Development Fund.

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, abs/1409.0473.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and

Marcello Federico. 2016. Neural versus Phrase-Based Machine Translation Quality: a Case Study. *CoRR*, abs/1608.04631.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany, August. Association for Computational Linguistics.

Castilho, Sheila and Sharon O'Brien. 2016. Evaluating the impact of light post-editing on usability. In *LREC*, pages 310–316, Portoroz, Slovenia, May.

Castilho, Sheila, Sharon O'Brien, Fabio Alves, and Morgan O'Brien. 2014. Does post-editing increase usability? A study with Brazilian Portuguese as Target Language. In *Proceedings of European Association for Machine Translation (EAMT)*, pages 183–190, Dubrovnik, Croatia.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017a. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108(1):109–120.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli Barone, and Maria Gialama. 2017b. A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In *MT Summit 2017*, Nagoya, Japan.

Castilho Monteiro de Sousa, Sheila. 2016. *Measuring acceptability of machine translated enterprise content*. Ph.D. thesis, Dublin City University.

Doherty, Stephen and Sharon O'Brien. 2012. A user-based usability assessment of raw machine translated technical instructions. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas*, pages 1–10, San Diego, California, USA.

Fuji, Masaru, N Hatanaka, E Ito, S Kamei, H Kumai, T Sukehiro, T Yoshimi, and Hitoshi Isahara. 2001. Evaluation method for determining groups of users who find mt useful. In *MT Summit VIII: Machine Translation in the Information Age*, pages 103–108.

Fuji, Masaru. 1999. Evaluation experiment for reading comprehension of machine translation outputs. In *Proceedings of MT Summit VII*, pages 285–289.

Jean, Sébastien, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal Neural Machine Translation Systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal, September.

Jones, Douglas, Edward Gibson, Wade Shen, Neil Granoien, Martha Herzog, Douglas Reynolds, and Clifford Weinstein. 2005. Measuring human readability of machine generated text: three case studies in speech recognition and machine translation. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 5, pages v–1009. IEEE.

Jordan-Nez, Kenneth, Mikel L Forcada, and Esteve Clua. 2017. Usefulness of mt output for comprehension an analysis from the point of view of linguistic intercomprehension. volume 1, pages 241–253, September.

Klerke, Sigrid, Sheila Castilho, Maria Barrett, and Anders Søgaard. 2015. Reading metrics for estimating task efficiency with mt output. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, pages 6–13.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. *CoRR*, abs/1706.03872.

Roturier, Johann. 2006. *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*. Ph.D. thesis, Dublin City University.

Scarton, Carolina and Lucia Specia. 2016. A reading comprehension corpus for machine translation evaluation. In Chair), Nicoletta Calzolari (Conference, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Stymne, Sara, Henrik Danielsson, Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz Lillkull, and Martin Wester. 2012. Eye tracking as a tool for machine translation error analysis. In *LREC*, pages 1121–1126.

Tomita, Masaru, Masako Shirai, Junya Tsutsumi, Miki Matsumura, and Yuki Yoshikawa. 1993. Evaluation of mt systems by toefl. In *Proceedings of the Theoretical and Methodological Implications of Machine Translation (TMI-93)*, pages 252–265.