Assessing the Usability of Raw Machine Translated Output: A User-Centred Study using Eye Tracking

Stephen Doherty
Centre for Next Generation Localisation
Centre for Translation and Textual Studies
Dublin City University
stephen.doherty@dcu.ie

Sharon O'Brien
Centre for Next Generation Localisation
Centre for Translation and Textual Studies
Dublin City University
sharon.obrien@dcu.ie

Abstract

This paper reports on the results of a project that aimed to investigate the usability of raw machine translated technical support documentation for a commercial online file storage service. Adopting a user-centred approach, we utilize the ISO/TR 16982 definition of usability - goal completion, satisfaction, effectiveness, and efficiency – and apply eye-tracking measures shown to be reliable indicators of cognitive effort, along with a post-task questionnaire. We investigated these measures for the original user documentation written in English and in four target languages: Spanish, French, German and Japanese, all of which were translated using a freely available online statistical machine translation engine. Using native speakers for each language, we found several significant differences between the source and MT output, a finding that indicates a difference in usability between well-formed content and raw

machine translated content. One target language in particular, Japanese, was found to have a considerably lower usability level when compared with the original English.

Keywords: usability, user-centred translation, eye tracking, machine translation, cognitive load.

1. Introduction

Machine Translation (MT), or automatic translation by computer, has enjoyed a considerable increase in utilization and popularity in the last decade or so, mainly thanks to a change in paradigm from a linguistic rules-based approach to a corpus-driven statistical approach (cf. Koehn, 2010). While these improvements in the underlying machine translation technology have led to better quality output (Hutchins, 2001; Callison-Burch *et al.,* 2008; Lopez, 2008; Specia *et al.,* 2009), outside of controlled or domain-specific contexts the output tends to require human intervention, known as *post-editing*, in order to meet the quality standards of professional human translation. There are many examples in the literature of the success of combining MT with a human post-editing process (e.g. Vasconcellos, 1985; Ørsnes *et al*., 1996; Senez, 1998; Groves, 2008; Roturier, 2009; O'Brien, 2006; Plitt and Masselot, 2010; Guerberof, 2012).

Notwithstanding the requirement for post-editing to meet high quality levels, the growth in popularity of free online MT systems has led to MT being increasingly used by non-specialists for a variety of purposes. Such users may, for instance, use an online MT engine to *gist*, i.e. to roughly understand the meaning of a text translated from a language they do not understand. While the goal of Fully Automatic High-

Quality Translation (FAHQT) remains unfulfilled, some companies and individuals adopt the approach of Fully Automatic Useful Translation (FAUT) for scenarios where only the gist of the meaning is required or very high quality publishable output is not required or, more commonly, cannot be paid for. Much recent research on MT output has focused on quality measurement of raw output (i.e. output that has not been post-edited by a human) and on the effort involved in post-editing. However, generally speaking, the usability of content is rarely considered from a translation perspective, let alone from a computer-aided translation perspective. Some limited examples exist, such as the discussion by Sacher *et al*. (2001) which suggests that the design of interactive products (especially for non-Western languages such as Chinese) should be tackled from a *languaculture* perspective and not from a 'translation problem' or 'deficit' perspective (ibid.: 43). Translation of content receives some, though very limited, attention in an article by Proctor *et al*. (2002), which considers generic problems concerning content preparation and management for web design. Here, translation of content is presented only as a 'problem for global e-commerce' and a 'difficult design challenge' (ibid: 70) and the authors warn that a site's information 'should not lose its meaning through translation' (ibid). It is apparent that usability of translated content remains relatively unexplored, and how *usable* raw MT output is for the end user and, indeed, how translation and its associated technology can be understood and explored from a HCI perspective (for a discussion, see O'Brien, 2012), needs further attention. By drawing attention to this topic, we aim to further interdisciplinary interaction between HCI and translation researchers, something that is identified as being important (Sears *et al.*, 2008; Karamanis *et al.,* 2009).

In the context of the current study, we propose to measure the usability of raw machine translated output for end users of an online file storage service. The question we propose to investigate is: are there significant differences in usability between the source language (SL) instructions (in English) and machine translated target languages (TL) as measured via screen recording, eye tracking and a post-task questionnaire?

While the quality of a system's output is an important factor of a user's satisfaction, acceptance, and performance, the user's perception and interaction with the system is also key (Hutchins, 2001), but is often overlooked as research on MT tends to focus on system development and evaluation (Reiter and Belz, 2009; Karamanis *et al*., 2009), and tends to ignore human factors (Dillinger and Lommel, 2004). Inclusion of users in evaluation of MT systems can provide benefits in both directions: such as positive influences on system development and its usability (Flournoy and Callison-Burch, 2001) which in turn lead to better systems and better output, making life easier for specialist users such as student and professional translators (O'Brien, 2006; Doherty *et al*., 2012).

While there are relatively few studies on the usability of raw machine translated output, traditional usability measures such as efficiency, accuracy and user satisfaction are of importance when assessing the usability of natural language processing applications, including online MT systems and their outputs (Dybkjaer *et al*., 2004). A general criticism of such usability studies is that they contain tasks that are designed by researchers but which may be meaningless for the actual users (Karamanis and Luz, 2009), something that may skew data and should be addressed appropriately in terms of materials and experiment design (see Section 2).

The current study builds upon previous work conducted by the researchers in which the usefulness of eye tracking as a semi-automatic method for the evaluation of MT output was investigated (Doherty and O'Brien, 2009; Doherty *et al.,* 2010, Doherty and O'Brien, 2012), in addition to the effects of controlled authoring[1] on the readability and comprehension of MT output (O'Brien, 2010; Doherty, 2012). Related work includes that of Tomita (1992) and Tomita *et al*. (1993), who examined the comprehension of MT output, and of Fuji (1999), who, in addition to this measure, included informativeness and fluency. Fuji *et al.* (2001) later focused on a measurement of usefulness as a combination of comprehension, time taken to answer questions relating to the text, and the user's subjective impression of the text which is described as the inverse of awkwardness. Usefulness is also examined by Roturier (2006) alongside comprehension and acceptability as rated by the user. Similarly, Jones *et al*. (2005) examine the readability of the output generated by an MT system by means of measuring accuracy in answering questions, task time, and a subjective rating given by users. Lastly, Gaspari (2004) evaluated users' needs when using online MT systems in terms of guessability (the user's expectations when first using an online system) and learnability (the capacity of the system to enable a user to learn how to use it), where the focus rested on the systems themselves. In the current study, we focus on MT output as the users in our study do not interact directly with the MT system.

A trend is observed in the above studies, and indeed in usability studies in general, where there is a divergence around the operational definition of usability, and, consequently, in its constituent measurements and conceptualisations. Here we

[1] Also known as controlled language and (linguistic) pre-processing. This involves authoring source text according to specific linguistic rules to make it less ambiguous and more suitable for MT.

adopt the ISO/TR 16982 definition of usability which is understood as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (International Organization for Standardization, 2002), which, in adhering to this ISO standard, we define here as:

i. Satisfaction: a measurement of user satisfaction of the instructions on a post-task 5-point Likert scale;
ii. Goal completion: a measurement of how successful the users are at accomplishing tasks, which were guided by the documented instructions as scored by the researchers using the gaze replay function of eye-tracking software;
iii. Total task time: a measurement of the overall duration of the tasks in seconds;
iv. Efficiency: measured as the number of successful tasks completed (out of all possible tasks) when total task time is taken into account, i.e.

$$\sum \frac{accuracy}{total\_task\_time(\text{sec.})} x100, where \frac{task\_sucesses}{total\_tasks} x100 = accuracy$$

We use these measures in conjunction with eye-tracking measures (see Sections 2 and 3 below) which, in addition to a post-task questionnaire, allow us to calculate the aforementioned usability measures and a variety of related measures, particularly those pertaining to cognitive effort.

2. Experiment Design

An integral part of the research project was to include a realistic task for plausible potential users of an online file storage service. The authors selected English documentation for the service and edited it to form a series of six paragraphs that comprised of one main task each and several specific subtasks, as follows:

i. Logging In (4 subtasks)
ii. Creating a New Folder (9 subtasks)
iii. File Management (6 subtasks)
iv. File Sharing (5 subtasks)
v. Folder Sharing (8 subtasks)
vi. Maintenance (7 subtasks)

As native speakers of English, the authors judged this documentation to be of good publishable quality and well-formed. Further to this, the online file storage service has a user base of over 50 million (Barret, 2011), a secondary indication at least that the instructions provided for the service are usable by the user base. These were initial assumptions that would be tested in the project. A non-domain specific freely available statistical machine translation system was used to translate the documentation (translate.google.com). This system was selected also because the scenario of a real end user making use of the output from this type of system (as opposed to a domain-specific engine or commercial/professional one) to translate documentation for comprehension purposes was realistic. The documentation was translated into French, German, Spanish, and Japanese as these were the languages for

which we could recruit native speakers as experimental participants and for which the service provides a somewhat localised version of the service.

Thirty participants were recruited and assigned to conditions based on their native language. The criteria for inclusion as participants were: (1) the participant was a native speaker of the source or target languages; (2) they had not yet used the online service but (3) they were a prospective user because they use computers and create electronic documents on a regular basis, which they may wish to store online or share and (4) they were willing to give consent to participate, on a voluntary basis, in a research project involving eye tracking and other measurements. Each participant viewed only the instructions in their native language, and no other versions of the text.

A Tobii 1750 eye tracker was used to record the reading of the instructions and execution of the required tasks. These instructions were placed on the left-hand side of the screen with the online storage service screen on the right-hand side (see Figure 1). We adhere to the guidelines for the presentation of onscreen stimuli as described in Gerganov (2007), who prescribes a screen resolution of 1280 by 1024 pixels, a font style of Tahoma, double line spacing, and a maximum of 90 characters per line. While a font size of 20 is recommended (ibid.), due to space constraints on the screen a font size of 16 was used. Finally, we use a threshold of 175ms as a minimum for average fixation duration which has been supported as the recommended value for reading tasks (Rayner and Sereno, 1994; Jensen *et al*., 2009). While the experiment was carried out in an eye-tracking lab, this lab is set up as an office and participants were made to feel that the task and environment were as naturalistic as possible. This set-up is akin to what Kjeldskov and Skov call ‘in sitro’ usability evaluations, which are ‘controlled, high fidelity simulations of the real world’ (2007: 31), which they found to be a reliable usability evaluation set-up.

Figure 1 Screen Setup for Experiment (Japanese)

3. Results

Section A describes participant information and results from the post-task questionnaire. Section B provides the eye-tracking data according to several measures: task time (observation length), fixation count, fixation duration, and attentional shifts. Lastly, Section C details the usability variables according to the ISO/TR 16982 definition. Each section deals with the variance between all languages, i.e. both the source and target languages. Initial findings from four of the usability measures used here, and without eye-tracking data, have been examined in terms of source language (English) versus target language groupings, where each of the target languages were treated as one group (see Doherty and O'Brien, 2012). This analysis sought to establish what differences existed between the non-machine translated and machine translated conditions in general. Here we expand on that analysis by examining all of these languages individually and also by using measures of cognitive effort and usability by means of eye tracking and a post-task questionnaire. Although analysis per target language grouping means reducing the number of participants in each group and, with that, the reliability of the results per target language, this was a logical next step after the larger initial group comparison. This focus on target language groupings allows us to identify trends in specific target languages, in turn allowing us to focus on those target languages that present the greatest usability challenges for machine translated content. Finally, given the diversity of results for

different language pairs (as is also evident in the current study), it is preferable to include as many languages as feasible in studies of MT systems and their output.

Section A: Participant Variables

3.1 Participant Information

Participants were categorised into language conditions, i.e. the version of the text that they read exclusively. The source condition (n = 15) contained the original English source text and was read by English native speakers only. The target conditions (n = 15) contained the machine-translated versions of the same instructions into French (4), German (3), Spanish (4), and Japanese (4) using the MT system. These conditions also included only native speakers for the respective language. Participants were instructed to read the instructions presented on the left side of the screen and to carry out the tasks as instructed.

3.2 Post-Task Questionnaire

The post-task questionnaire was designed to measure self-report for items pertaining to several aspects of usability, which included those from the ISO definition (satisfaction), and additional factors such as comprehension, and recommendation. The questionnaire contained 12 items, each of which asked participants to report the extent to which they agreed or disagreed (on a 5-point Likert scale) with a statement relating to the instructions they had just read. Additional

demographic data and details pertaining to professional experience were also collected in these items, as well as one item testing memory recall.

I. Comprehension

For the first item "the instructions were comprehensible", a one-way ANOVA found a significant difference between languages, where F (4, 25) = 6.41, p = .001. Tukey post-hoc comparisons of the five groups indicate that the English group gave significantly higher ratings for comprehension than the French (p = .036) and Japanese groups (p = .003) - the differences between the other groups were not significant - see Table 1. In other words, the differences between English-German and English-Spanish were not significant[2], and no significant differences were observed between each of the other language interactions: French-German, French-Spanish, French-Japanese, German-Spanish, German-Japanese; and Spanish-Japanese.

	Mean	SD	CIs (95%) Lower	Upper
English	4.53	.516	4.25	4.82
French	3.00	1.41	0.75	5.25
German	3.00	1.00	0.52	5.48
Spanish	4.25	0.96	2.73	5.77
Japanese	2.50	1.29	0.45	4.55

Table 1 Comprehension Ratings for Each Language

II. Task Completion

[2] While the means of the French and German groups are the same, this difference was not significant compared to English due to the smaller sample size of the German group.

For the second questionnaire item "I could complete the task by following the instructions provided", a one-way ANOVA found a significant difference between languages, where F (4, 25) = 6.117, p = .001. Tukey post-hoc comparisons of the five groups indicate that the English group gave significantly higher ratings for this measure than the Japanese group (p < .05). The other groups did not differ significantly in their interactions with each other - see Table 2.

	Mean	SD	CIs (95%) Lower	Upper
English	4.53	0.64	4.18	4.89
French	3.75	0.50	2.95	4.55
German	3.67	0.58	2.23	5.10
Spanish	4.00	1.41	1.75	6.25
Japanese	2.50	0.58	1.58	3.42

Table 2 Ratings for Instructions Allowing Task Completion for Each Language

III. Satisfaction

For the third item "I was satisfied with the instructions provided", a one-way ANOVA found a significant difference between languages, where F (4, 25) = 8.341, p < .001. Tukey post-hoc comparisons of the five groups indicate that the English group gave significantly higher ratings for satisfaction than the French (p < .05), German (p < .05), and Japanese (p < .05) groups; the other groups did not differ significantly in their interactions with each other – see Table 3.

	Mean	SD	CIs (95%) Lower	Upper
English	4.13	0.74	3.72	4.54
French	2.25	1.50	-0.14	4.64
German	1.33	0.58	-0.10	2.77
Spanish	3.25	2.06	-0.30	6.53
Japanese	1.50	0.58	0.58	2.42

Table 3 Satisfaction Ratings for Each Language

IV. Potential Improvement

For the fourth item "the instructions could be improved upon", a one-way ANOVA found no significant difference between languages at ($p < .05$), i.e. participants indicated that each of the versions of the instructions, even including the source language English, could be improved upon. The English and Spanish versions were given the lowest of all ratings for potential improvement, but had an average score of approximately 3.5 on the 5-point scale – see Table 4. In other words, these two versions required the least amount of improvement, but the users' opinions were that the instructions could nevertheless be ameliorated.

	Mean	SD	CIs (95%) Lower	Upper
English	3.40	1.40	2.62	4.18
French	5.00	0.00	5.00	5.00
German	5.00	0.00	5.00	5.00
Spanish	3.50	1.73	0.74	6.26
Japanese	5.00	0.00	5.00	5.00

Table 4 Ratings for Potential Improvement for Each Language

V. Future Reuse

For the fifth item "I would be able to use the software again in the future without the instructions", a one-way ANOVA found no significant difference between languages at ($p > .05$) – see Table 5. This result indicates that participants in each group reported that they would not require the instructions for future use of the online storage service and suggests that, despite quality issues, the instructions were good

enough to allow users to acquire knowledge of how to use the service, or it could even suggest that the instructions were superfluous to completing the tasks assigned.

	Mean	SD	CIs (95%) Lower	Upper
English	4.20	1.01	3.64	4.76
French	4.25	0.96	2.73	5.77
German	3.33	0.58	1.90	4.77
Spanish	4.00	0.82	2.70	5.30
Japanese	3.25	0.50	2.45	4.05

Table 5 Ratings for Reuse without Instructions for Each Language

VI. Recommendation

For the sixth item "I would recommend the software to a friend/colleague", a one-way ANOVA found a significant difference between languages, where F (4, 25) = 6.195, p = .001. Tukey post-hoc comparisons of the five groups indicate that the English group gave significantly higher ratings for recommendation than the German ($p < .05$) and Japanese ($p < .05$) groups. The other groups did not differ significantly from each other in their interactions - see Table 6.

	Mean	SD	CIs (95%) Lower	Upper
English	4.40	0.63	4.05	4.75
French	3.50	1.29	1.45	5.55
German	2.33	0.58	0.90	3.77
Spanish	3.75	0.96	2.23	5.27
Japanese	2.50	1.29	0.45	4.55

Table 6 Ratings for Recommendation for Each Language

VII. Recall

Finally, for the recall test item where participants were asked how much file space was being used in the tool (information they should have read in their instructions), a one-way ANOVA found a significant difference between languages, where F (4, 25) = 7.147, p = .001. For this measure, participants scored a 1 for a correct answer, and a 0 for an incorrect or incomplete answer. Tukey post-hoc comparisons of the five groups indicate that the English group were significantly more successful for recall than the German ($p < .05$) and Japanese ($p < .05$) groups – see Table 7. The other groups did not differ significantly from each other in their interactions. This indicates that for cued recall (where participants were not forewarned about explicit recall testing) the English version was more likely to result in retention, despite participants in all groups having to view the dialog box stating how much file space was being used in order to log out of the service. A plausible explanation is that the better quality of the English version allowed participants to spend more of their task time in the task window where they may have noticed more about the online service's environment – we return to this is in Section 3.8.

	Mean	SD	CIs (95%) Lower	Upper
English	0.87	0.35	0.67	1.06
French	0.50	0.58	-.042	1.42
German	0	0	0	0
Spanish	0.25	0.50	-0.55	1.05
Japanese	0	0	0	0

Table 7 Recall Test Scores for Each Language

3.3 Correlational Analysis

A correlational analysis (see Table 8) showed a strong positive correlation between all but two of the interactions, namely: reuse and potential improvement, and

recall and potential improvement. While these were both moderate correlations, neither was significant. The results for the other interactions demonstrate the construct validity of the measures; for example, ratings for comprehension were a strong positive correlate of satisfaction and strong negative correlate of ratings for potential improvements.

	Comp.	Complete Tasks	Sat.	Could Be Improved	Reuse	Rec.	Recall
Comprehensible	-	.744**	.842**	-.702**	.436*	.611**	.481**
Complete Tasks	.744**	-	.791**	-.648**	.595**	.706**	.456*
Satisfaction	.842**	.791**	-	-.814**	.570**	.757**	.457*
Could Be Improved	-.702**	-.648**	-.814**	-	-.355	-.627**	-.339
Reuse	.436*	.595**	.570**	-.355	-	.573**	.569**
Recommend	.611**	.706**	.757**	-.627**	.573**	-	.545**
Recall	.481**	.456*	.457*	-.339	.569**	.545**	-

Table 8 Correlation Coefficients for Questionnaire Variables (ρ)[3]

Overall, the results indicate that the English group rated the instructions consistently higher than each of the other groups across the measures on the post-task questionnaire. Significant differences were found for: comprehension ratings of the instructions, the value of the instructions for task completion, satisfaction with the instructions, likelihood to recommend the software/file storage service to others, and recall. Among the significant differences, Japanese was shown to perform worst of all (5 instances of significantly lower score), followed by German (3 instances), and French (2 instances). No significant differences were observed for the measures of reuse of the tool without the instructions, or for the need to improve the instructions. In other words, participants felt that the instructions would not be required for future

[3] * Correlation is significant at the .01 level (2-tailed); ** correlation is significant at the .05 level (2-tailed).

use of the tool, perhaps due to the simplicity of the introductory tasks, rather than due the quality of the instructions. Finally, while there are no significant differences between English and Spanish for any of the above measures, Spanish has lower scores for all variables.

Section B: Eye-Tracking Variables

3.4 Section Overview

This section examines the results from the eye-tracking variables for English and each of the target languages for the entire experiment exclusive of the post-task questionnaire which was administered as a hard copy. We first examine total task time, and then move to total fixation count and average fixation duration. Fixations are defined as "eye movements which stabilize the retina over a stationary object of interest" (Duchowski 2003, p. 43) and occur when the eye focuses on a particular item, e.g. a word on the screen. Longer task times, higher fixation counts and their average durations have been shown to be reliable indicators of cognitive effort across a variety of tasks (cf. Rayner, 1998; Schultheis and Jameson, 2004; Iqbal *et al*., 2005; Stanford Poynter Project).

Following this, the aspect of attentional switching between the instructions window (where the instructions were presented) and the task window (where the participants interact with the online file storage service) is discussed with a presentation of the number of attentional shifts between these windows. Attention refers to the ability of the working memory to hold into focus one task or object in particular to the exclusion or filtering out of other stimuli due to its limited processing

resources (Baddeley, 2007). While attention can be divided between tasks (ibid.), the switching of attention from one task to another incurs a cost in terms of cognitive resources, and has been found to result in increased processing times (Baddeley, 2007; Hvelplund, 2011) and increased fixations (Jakobsen and Jensen, 2008).

3.5 Total Task Time

Task time is recorded in milliseconds but reported here in seconds. It denotes the time from the activation of the eye-tracking recording to the completion of the tasks, where the participants signalled to the researcher that they had completed their tasks. A one-way ANOVA found a significant difference between languages, where F (4, 24) = 2.814, p = .048. Tukey post-hoc comparisons of the five groups indicate that the English group had significantly shorter task times overall than Japanese ($p < .05$). The other groups did not differ significantly from each other in their interactions - see Table 9.

	Mean	SD	CIs (95%) Lower	Upper
English	371.9	183.6	265.8	477.9
French	563.5	471.9	-187.4	1314.5
German	883.9	448.2	-279.5	1947.3
Spanish	489.2	307.8	-0.7	979.0
Japanese	878.0	447.3	166.2	1589.8

Table 9 Task Times (in seconds) for Each Language

3.6 Fixation Count

With regard to the total number of fixations, a one-way ANOVA found a significant difference between languages, where F (4, 24) = 3.293, p = .028. Tukey

post-hoc comparisons of the five groups indicate that the English group had significantly fewer fixations than the Japanese group ($p < .05$). The other groups did not differ significantly from each other in their interactions - see Table 10.

	Mean	SD	CIs (95%) Lower	Upper
English	646.3	367.6	434.1	858.5
French	752.0	195.9	400.2	1063.9
German	1040.0	231.5	464.9	1615.1
Spanish	1225.5	783.0	-20.5	2471.5
Japanese	1467.9	633.4	459.2	2474.8

Table 10 Fixation Count for Each Language

3.7 Average Fixation Duration

A one-way ANOVA found a significant difference between languages, where F (4, 24) = 2.695, p = .05. Tukey post-hoc comparisons of the five groups indicate that the English group had shorter average fixation durations than the Japanese group ($p < .05$). The other groups did not differ significantly from each other - see Table 11.

	Mean	SD	CIs (95%) Lower	Upper
English	152.6	98.9	95.5	209.7
French	187.7	94.8	36.8	338.7
German	316.2	143.2	-39.6	672.1
Spanish	214.1	66.2	49.7	378.4
Japanese	333.5	183.3	41.9	625.1

Table 11 Mean Fixation Duration (in milliseconds) for Each Language

3.8 Attentional Shifts

In our investigation of participants' attention during the experiment, we first examine the foci participants had on the instructions window (where the instructions

were displayed) and the task window (where the online file storage service could be used). This was carried out by means of isolating AOIs around these windows and capturing the number and durations of fixations for each window. Following on from this, we count the number of shifts, i.e. where a fixation lies in one AOI and the next fixation occurs in the other AOI, indicating an expenditure of cognitive effort in the form of processing times.

For the amount of time spent in each window, a one-way ANOVA found no significant difference between languages, where F (4, 25) = 1.793, p = .162. While the English group (see Table 12) spent a greater percentage of their time on the task window than the other languages, Japanese participants spent the greater amount of their time on the instructions window - an indication of a greater need to attend to the instructions located in this part of the screen.

	Mean	SD	CIs (95%) Lower	Upper
English	45.4	7.4	41.3	49.6
French	50.3	13.2	29.3	71.3
German	51.2	14.6	14.9	87.5
Spanish	52.9	13.2	31.9	73.8
Japanese	59.1	5.7	49.9	68.2

Table 12 Percentage of Time Spent in Instructions Window

With regard to the shifting of attention, a one-way ANOVA found a significant difference between languages, where F (4, 25) = 4.497, p = .007. Tukey post-hoc comparisons of the five groups indicate that the English group had significantly fewer shifts of attention between windows than the Japanese group (p < .05), while the other groups did not differ significantly from each other - see Table 13.

	Mean	SD	CIs (95%) Lower	Upper

English	54.0	10.3	48.3	59.7
French	62.3	6.7	51.7	72.8
German	61.0	3.6	52.0	69.9
Spanish	58.0	4.2	51.3	64.8
Japanese	73.8	6.4	63.6	83.9

Table 13 Number of Attentional Shifts between Windows

3.10 Correlational Analysis

A correlational analysis (see Table 14) showed a series of strong positive correlations for all eye-tracking variables.

	Task Time	**Fixation Count**	**Average Fixation Duration**	**Attentional Shifts**
Task Time	-	.677**	.687**	.573**
Fixation Count	.677**	-	.824**	.504**
Average Fixation Duration	.697**	.824**	-	.463**
Attentional Shifts	.573**	.504**	.463*	-

Table 14 Correlation Coefficients for Eye-Tracking Variables (*r*)

Overall, the results from the eye-tracking measures show that task time was shorter for English (significantly so when compared to the Japanese group). The total fixation count and average durations were also lower for the English-speaking users (significantly against the Japanese group for both measures). With regard to the percentage of time the users spent between the two windows, no significant difference was found. However, the Japanese group spent more of their time on the instructions window (59%) while the English group spent a greater percentage of their time using the file storage service in the task window (45%) when compared with the other groups. In terms of switching attention between the windows, the English group did so much less frequently than the other groups (significant against Japanese). Once

again we see a trend towards the English group performing better than the others, closely followed by Spanish, and with Japanese consistently lower in terms of usability variables resulting in longer task times, more fixating, and rereading.

Section C: Usability Variables

This section presents the variables used to measure usability. As described previously in Section 1, by using the ISO/TR 16982 ISO definition for usability, we examine: satisfaction; goal completion); total task time; and efficiency.

3.11 User Satisfaction

As described in Section A of the results, a significant difference was found between languages [F (4, 25) = 8.341, $p < .001$], where English was rated to be more satisfactory than the other language versions, significantly so for French, German, and Japanese, as seen in Table 3.

3.12 Goal Completion

Overall, the scores on goal completion for each language were quite high, i.e. participants were largely successful across all language conditions, especially for English and Spanish. Out of the maximum possible score of 30 for the experiment (where each task was assigned one or zero points based on its completion or lack thereof), a one-way ANOVA found a significant difference between languages, where F (4, 25) = 5.822, p = .02. Tukey post-hoc comparisons of the five groups indicate

that the English and Spanish groups were significantly more successful at completing goals than the Japanese group ($p < .001$). However, the other groups did not differ significantly - see Table 15.

	Mean	SD	CIs (95%) Lower	Upper
English	30.0	0	30.0	30.0
French	28.0	1.6	25.4	30.6
German	27.3	4.6	15.9	38.9
Spanish	30.0	0	30.0	30.0
Japanese	25.5	3.4	20.0	30.9

Table 15 Scores for Goal Completion for Each Language (Max. = 30)

3.13 Total Task Time

As described in Section B, a one-way ANOVA found a significant difference between languages, where F (4, 24) = 2.814, p = .048. Tukey post-hoc comparisons of the five groups indicate that the English group had significantly longer task times than the Japanese ($p < .05$) group, while other groups did not differ significantly from each other – as seen in Table 9.

3.14 Efficiency

As described in Section 1, efficiency is calculated by the number of successful tasks completed against the total number of tasks in the experiment, then expressing this result as a divisor of the respective total task time. A one-way ANOVA found a significant difference between languages, where F (4, 25) = 4.87, p = .005. Tukey post-hoc comparisons of the five groups indicate that the English language group was significantly more efficient than the Japanese ($p < .05$) and German groups ($p < .05$),

while the other groups did not differ significantly - see Table 16. It should be noted that higher scores for the efficiency variable indicate greater efficiency.

	Mean	SD	CIs (95%) Lower	Upper
English	39.6	13.1	32.4	46.8
French	21.3	11.4	3.0	39.5
German	16.3	11.9	-13.3	45.9
Spanish	27.3	10.8	10.1	44.4
Japanese	17.8	10.4	1.1	34.4

Table 16 Efficiency Scores for Each Language

3.15 Correlational Analysis

A correlational analysis of the usability variables (see Table 17) showed strong positive correlations between all but one of the interactions: task time and goal completion. It stands to reason that user satisfaction correlates with efficiency in that participants were aware of their completion of tasks with relative ease and encountered few, if any, problems resulting in a satisfactory user experience. Once again, as efficacy is a derivative of task time, the strong relationship between these two variables is to be expected. Lastly, as goal completion was an arbitrary and constructed value (a continuous variable) for this experiment, i.e. 30 tasks that were completed in the majority of cases, it is unsurprising that it does not correlate as strongly, especially with task time (a discrete variable).

	Satisfaction	Goal Completion	Task Time	Efficiency
Satisfaction	-	.629**	-.548**	.522**
Goal Completion	.629**	-	-.344	.479**
Task Time	-.548**	-.344	-	.744**
Efficiency	.522**	.479**	.744**	-

Table 17 Correlation Coefficients for Usability Variables (ρ)

Overall, from the results of the usability measures, we see that the English users were more satisfied than the other groups (significant against Japanese). For the completion of goals, the English and Spanish groups outperformed the other groups, significantly so when compared with Japanese. As reported above, task time was shorter for English and, for the measure of efficiency, the English users were significantly more efficient than both the Japanese and German users, closely followed by the Spanish users.

4. Conclusions

The aim of the current study was to ascertain the usability of raw machine translated text as translated by a freely available online MT system for end users of a popular online file storage service. Our stated research question was: are there significant differences in usability between the source language (SL) instructions (English) and machine translated target languages (TL) as measured via screen recording, eye tracking and a post-task questionnaire?

Firstly, from the results of the post-task questionnaire we can conclude that the English source had the highest ratings for comprehension, satisfaction, recommendation to others, and recall. While the instructions were reported to enable users to carry out the required tasks, all language versions of the instructions were reported to have room for potential improvement and were deemed unnecessary for using the file storage service again in the future.

Secondly, in terms of the results from the eye-tracking measures, we can conclude that processing the English instructions required less cognitive effort as

measured by fixation count, duration and overall task time. Time spent between task and instructions windows on screen was largely uniform; however, shifts of attention between them were much less prevalent for the English group. Thirdly, in examining the results of the other usability measures we conclude that the English and Spanish instructions score significantly higher for goal completion than the Japanese group, and that the English group was more efficient than the others, significantly so compared with German and Japanese groups.

As evident from the series of results summarised here and presented in more detail in the previous sections, there is an overall consensus between the different measures employed in the study that raw MT output is less usable than the source language documentation written in English. For the most part, these measures correlate with each other and yield consistent results.

Discussion

This study has sought to fill the gap in research on machine translation and usability. We conclude that the raw MT output used in this experiment has a reasonable level of usability and certainly gave our participants more than just a *gist* of the meaning. However, in almost all measures, the source language instructions score higher in our usability measurements than the machine translated target languages. The Spanish target language was notably higher scoring than the others, while Japanese was notably lower scoring. This is not a surprising result; researchers in machine translation are aware that certain language pairs are considered to be reasonably 'easy' to machine translate, while others are more 'difficult'.

While the results show that raw MT output is indeed usable in real-world scenarios, they also demonstrate the added value of text produced by native speakers. If post-editing is considered to be a human intervention that raises raw MT output from the status of 'machine-generated' to 'native-like quality', then it seems that there is added value in post-editing for organisations who are concerned with user satisfaction.

In the case of the user scenarios presented here, it is evident that for certain languages the use of an online MT system is a viable solution to the service vendor (the online file storage service). However, while users of the tasks can indeed *use* the software (albeit less efficiently), their opinion and feelings towards the instructions, and consequently the overall user experience, suffers somewhat, especially for the French, German, and Japanese users. We do not rule out the potential effect of cultural differences here, in particular differences regarding the concept of language quality. However, a cross-cultural analysis of the perception of document quality was beyond the scope of this experiment. It would certainly be interesting to draw on cross-cultural usability research in the future to investigate this further (e.g. Tractinsky 1997; Herman 1996; Hall *et al.* 2004).

While it stands to reason, it is nevertheless important to report that users who comprehended the instructions were significantly more satisfied and therefore reported that they would be more willing to recommend the product to others in the future. This highlights the commercial ramifications of ensuring the user is presented with documentation that they can understand. Appropriate measures of comprehension should be used as they represent a more robust measure than simply asking a user to rate their subjective and often superficial understanding on a scale.

Limitations & Future Work

While the sample size is a limitation of the current study, the small number of participants is not unusual for eye-tracking studies of usability. The difficulty in accessing native speakers of the four target languages who were prepared to volunteer was one restriction. While the number is small for each target language, we have tried to compensate for this by also analysing the data according to two groups, source and target language (cf. Doherty and O'Brien 2012). At the same time, looking at results on an individual target language basis provides important information about differences across languages. This is important because positive or negative results for one language pair will not necessarily apply to other language pairs. Additionally, although the numbers are small, the 'in sitro' set-up (Kjeldskov and Skov, 2007) lends to the study's credibility. The study would benefit from increasing the number of participants and it is our intention to build on the research into usability of machine translated content in the future.

As stated earlier: there was a possibility that the instructions were not required to complete the tasks. A hypothesis that emerges is that users, even those new to the interface, could successfully complete the tasks if given a task list and no instructions by using prior experience as computer users. As a next step, we intend to use a 'control' group of users of the file storage service in order to examine the user experience when no written instructions are present. This will allow us further insight into the role instructions play in the first instance. In addition to this, a closer examination of affordances (see Gaver, 1991) would likely provide more evidence for the compensation the user, and indeed the system or UI, makes for instructions that are linguistically poor in quality.

In terms of the language used, we would like to expand to other language pairs given the great difference in MT quality for different languages (Koehn, 2005; Avramidis and Koehn, 2008). Moreover, it would be interesting to compare the usability of human-translated instructions against the raw machine translated instructions, or against machine translated, human post-edited instructions.

We have used only one, freely available MT engine in this experiment. A cross-engine comparison would also be useful to ascertain what effect different MT engines may have on the results. It has been shown that domain-specific engines perform better when trained on in-domain data (e.g. Haque *et al.*, 2009; Banerjee *et al.*, 2010; Núria *et al.,* 2012) as opposed to the large body of mixed data contained in the corpora used by freely available online engines, which are trained on large, general corpora whose quality cannot be guaranteed. A comparison across these engine types would also be useful in measuring relative usability according to engine type.

As mentioned above, the links between linguistic and cultural norms and expectations should not be excluded from usability studies of translated output, especially when comparing groups of language communities with each other, and thus we would encourage the inclusion of cross-cultural factors in future studies. By using more user-designed tasks, we hope to include, to some extent, a more diverse approach to intra- and intercultural differences that have been documented in the literature (Beu *et al*., 2000; Nisbett and Miyamoto, 2005; Dong and Lee, 2008).

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University. The authors would also like to express their gratitude to the participants of the study.

References

Avramidis, E., & Koehn, P. 2008. Enriching morphologically poor languages for statistical machine translation. *Proceedings of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, June 15-20, 2008, the Ohio State University, Columbus, Ohio, USA, 763-770.

Baddeley, A. 2007. *Working memory, thought, and action*. Oxford: Oxford University Press.

Banerjee, P., Du, J., Naskar, S. K., Li, B., Way, A. & van Genabith, J. 2010. Combining multi-domain statistical machine translation models using automatic classifiers. *Proceedings of AMTA 2010: the ninth conference of the Association for Machine Translation in the Americas*, Denver, Colorado, October 31 – November, 1-10.

Barret, V. (2011). Dropbox: The inside story of tech's hottest startup. *Forbes Magazine*, Nov., 7th, 2011. (Accessed: August 1st, 2012).

Beu, A., Honold, P. & Yuan, X. 2000. How to build up an infrastructure for intercultural usability engineering. *International Journal of Human-Computer Interaction*, 12(2), 347-358.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. 2008. Further meta-evaluation of machine translation. *Proceedings of ACL-08: HLT. Third*

Workshop on Statistical Machine Translation, June 19, 2008, the Ohio State University, Columbus, Ohio, U.S.A., 70-106.

Dillinger M., & Lommel, A. (2004). LISA best practice guide on MT. *LISA Case Studies.* Localization Industry Standards Association, Available from: http://www.translationoptimization.com/papers/DillingerLommel_MT_BPG.pdf

Doherty, S., & O'Brien, S. 2009. Can MT output be evaluated through eye tracking? *Proceedings of the MT Summit XII*, Ottawa, Ontario, Canada, 214-221.

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

Doherty, S. 2012. *Investigating the effects of controlled language on the reading and comprehension of machine translated texts: A mixed-methods approach*. Unpublished PhD thesis, Dublin City University.

Doherty, S., Kenny, D., & Way, A. 2012. Taking statistical machine translation to the student translator. *Proceedings of AMTA 2012: the tenth conference of the Association for Machine Translations in the Americas*, San Diego, California, U.S.A., no page numbers.

Doherty, S. & O'Brien, S. 2012. A user-based usability assessment of raw machine translated technical instructions. *Proceedings of AMTA 2012: the tenth conference of the Association for Machine Translations in the Americas*, San Diego, California, U.S.A., no page numbers.

Dong, Y., & Lee, K. P. 2008. A cross-cultural comparative study of users' perceptions of a webpage: With a focus on the cognitive styles of Chinese, Koreans and Americans. *International Journal of Design*, 2(2), 19-30.

Duchowski, A. T. 2003. *Eye tracking methodology: Theory and practice.* London: Springer-Verlag.

Dybkjær, L., Bernsen, N., & Minker, W. 2004. Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication,* 43(1-2), 33-54.

Flournoy, R. S., & Callison-Burch, C. 2001. Secondary benefits of feedback and user interaction in machine translation tools. *Proceedings of MT Summit VIII – MT2012 Towards a Road Map for MT*, Santiago de Compostela, Spain, no page numbers.

Fuji, M. 1999. Evaluation experiment for reading comprehension of machine translation outputs. *Proceedings of MT Summit VII*, Singapore, 285-289.

Fuji, M., Hatanaka, E., Ito, S., Kamai, H., Sukehiro, T., Yoshimi, T., & Ishara, H. 2001. Evaluation method for determining groups of users who find MT useful. *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain, 103-108.

Gaspari, F. 2004. Online MT services and real users' needs: An empirical usability evaluation. *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas*, Washington D.C., U.S.A., 74-85.

Gaver, W. 1991. Technology affordances. *Proceedings of CHI 1991*, ACM Press: New York, 79 – 84.

Gerganov, A. (2007). *Eye tracking studies with Tobii 1750 - Recommended settings and tests.* Available from: http://cogs.nbu.bg/eye-to-it/del/EYE-TOIT_D1.2_A.pdf, Last accessed: 20/03/13.

Groves, D. 2008. Bringing humans into the loop: localization with machine translation at Traslán. *Proceedings of AMTA 2008: the eighth conference of*

the Association for Machine Translation in the Americas, Waikiki, Hawai'i, 11-22.

Guerberof, A. 2012. *Productivity and quality in the post-editing of outputs from translation memories and machine translation*. Unpublished PhD thesis. Universitat Rovira I Virgili, Tarragona. Spain.

Hall, M., De Jong, M., & Steehouder, M. 2004. Cultural differences and usability evaluation: Individualistic and collectivistic participants compared. *Technical Communication*, 51(4), 489-503.

Haque, R., Naskar, S.K., van Genabith, J., & Way, A. 2009. Experiments on domain adaptation for English-Hindi SMT. *Proceedings of PACLIC 23: the 23rd Pacific Asia Conference on Language, Information and Computation Hong Kong*, 670–677.

Herman, L. 1996. Towards effective usability evaluation in Asia: Cross-cultural differences. *Proceedings of the Sixth Australian Conference on Computer-Human Interaction*, 135-136.

Hutchins, J. 2001. Machine translation and human translation: in competition or in complementation? *International Journal of Translation, 13* (1-2), 5-20.

Hvelplund, K. T. 2011. *Allocation of cognitive resources in translation: An eye-tracking and key-logging study*. Unpublished PhD thesis. Copenhagen Business School.

International Organization for Standardization. 2002. ISO/TR 16982: Ergonomics of human-system interaction – Usability methods supporting human centred design.

Iqbal, S., Adamzyck, P., Zheng, X., & Bailey, P. 2005. Towards an index of opportunity: understanding changes in mental workload during task execution.

Proceedings of CHI 2005: Human Factors in Computing Systems, ACM Press, New York, 311–320.

Jakobsen, A. L., & Jensen, K. T. H. 2008. "Eye movement behaviour across four different types of reading task". In Copenhagen Studies in Language 36: *Looking at Eyes: Eye-Tracking Studies of Reading and Translation Processing*, Edited by: Göpferich, S., Jakobsen, A. L. & Mees, I. M. 103-124. Copenhagen: Samfundslitteratur.

Jones, D., Gibson, E., Shen, W., Granoien, N., Herzog, M., Reynolds, D., & Weinstein, C. 2005. Measuring human readability of machine generated text: three case studies in speech recognition and machine translation. *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing,* Philadelphia, U.S.A., 1009-1012.

Jensen, K. T. H., Sjørup, A. C., Balling, L. W. 2009. "Effects of L1 syntax on L2 translation". In *Copenhagen Studies in Language 38: Methodology, Technology and Innovation in Translation Process Research*, Edited by: Mees I. M., Alves, F. and Göpferich, S. 319-336, Copenhagen: Samfundslitteratur.

Karamanis, N., & Luz, S. 2009. Interaction strategies by a non-English speaker in Dublin and their relation to machine translation. *Proceedings of the 3rd Irish Human Computer Interaction Conference*, Dublin, Ireland, 107-110.

Karamanis, N., Schneider, A., van der Sluis, I., Schlogl, S., Doherty, G., & Luz, S. 2009. Do HCI and NLP interact? *Proceedings of the 27th International Conference on Human Factors in Computing Systems, ACM Extended Abstracts,* 4333-4338.

Kjeldskov, J., & Skov, M. 2007. Studying usability in sitro: Simulating real world phenomena in controlled environments. *International Journal of Human-Computer Interaction*, 22(1-2), 7-36.

Koehn, P. 2005. Europarl: a parallel corpus for statistical machine translation. *Proceedings of MT Summit X*, Phuket, Thailand, 79-86.

Koehn, P. 2010. *Statistical machine translation*. London: Cambridge University Press.

Lopez, A. 2008. Statistical machine translation. *AMC Computing Surveys*, 40(3), 8:1–8:49.

Nisbett, R., & Miyamoto, Y. 2005. The influence of culture: holistic versus analytic perception. *Trends in Cognitive Science*, 9(10), 467-472.

Núria, B., Papavasiliou, V., Prokopidis, P., Toral, A., & Arranz, V. 2012. Mining and exploiting domain-specific corpora in the PANACEA platform. *Proceedings of the 5th Workshop on Building and Using Comparable Corpora: "Language Resources for Machine Translation in Less-Resourced Languages and Domains"*, LREC 2012 Workshop, Istanbul, Turkey, 24-26.

O'Brien, S. 2006. Machine-translatability and post-editing effort: an empirical study using Translog and Choice Network Analysis, Unpublished PhD thesis, Dublin City University.

O'Brien, S. 2010. "Controlled language and readability". In *Translation and Cognition: American Translators Association Monograph Series XV,* Edited by: Shreve, G. M. and Angelone, E. 143-165. Philadelphia: John Benjamins.

O'Brien, S. 2012. Translation as human-computer interaction. *Translation Spaces*, 1, 101-122.

Ørsnes, B., Bradley Music, B., & Maegaard, B. 1996. PaTrans – a patent translation system. *Proceedings of COLING 1996: the 16th International Conference on Computational Linguistics*, August 5-9, 1996, Center for Sprogteknologi, Copenhagen, 1115-1118.

Plitt, M., & Masselot, F. 2010. A productivity test of statistical machine translation post-editing in a typical localization context. *The Prague Bulletin of Mathematical Linguistics*, 23, 7-16.

Proctor, R., Vu, K. & Salvendy, G. 2002. Content preparation and management for web design: Eliciting, structuring, searching, and displaying information. *International Journal of Human-Computer Interaction* 14(1), 25-92.

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

Rayner, K., & Sereno, S. 1994. "Eye movements in reading: psycholinguistic studies". In *Handbook of Psycholinguistics*, Edited by: Gernsbacher M.A. 57-81. New York: Academic Press.

Reiter, E., & Belz, A. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics, 35*(4), 529-558.

Roturier, J. 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.* Unpublished PhD thesis, Dublin City University.

Roturier, J. 2009. Deploying novel MT technology to raise the bar for quality: a review of key advantages and challenges. *Proceedings of MT Summit XII*, Ottawa, Canada, 1-8.

Sacher, H., Tng, T., & Loudon, G. 2001. Beyond translation: Approaches to interactive products for Chinese consumers. *International Journal of Human-Computer Interaction,* 13(1), 41-51.

Schultheis, H., & Jameson, A. 2004. "Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioural methods". In *Adaptive Hypermedia and Adaptive Web-based Systems,* Edited by: Neijdl W & de Bra, P. 18-24. Eindhoven: Springer Verlag.

Sears, A., Lazar, J., Ozok, A., & Meiselwitz, G. 2008. Human-centered computing: Defining a research agenda. *International Journal of Human-Computer Interaction* , 24(1), 2-16.

Senez, D. 1998. The machine translation help desk and the post-editing service. *Terminologie et Traduction, 1*, 289-295.

Specia, L., Turchi, M., Wang, Z., Shawe-Taylor, J., & Craig Saunders, C. 2009. Improving the confidence of machine translation quality estimates. *Proceedings of MT Summit* XII, Ottawa, Canada, 136-143.

Stanford Poynter Project: http://www.poynterextra.org/et/i.htm, Last accessed: 09/08/2012.

Tomita, M. 1992. Application of the TOEFL test to the evaluation of Japanese-English MT. *Proceedings of MT Evaluation Workshop, AAMT*, November, 1992, 59-60.

Tomita, M., Shiri, M., Tsutsumi, J., Matsumura, M., & Yoshikawa, Y. 1993. Evaluation of MT systems by TOEFL. *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation*. 252-265.

Tractinsky, N. 1997. Aesthetics and apparent usability: empirically assessing cultural and methodological issues. *Proceedings of the SIGCHI conference on human factors in computing systems*. Atlanta, GA, USA, 115-122.

Vasconcellos, M. 1985. Management of the machine translation environment: Interaction of functions at the Pan American Health Organization. *Proceedings of Translating and the Computer 5: Tools of the Trade*. London: Aslib, 115-129.

Author Biographies:

Dr. Stephen Doherty is a post-doctoral researcher in the areas of language, cognition, and human-computer interaction. He is currently a post-doctoral research fellow funded by Science Foundation Ireland and based in the School of Applied Language and Intercultural Studies in Dublin City University, where he lectures in translation technologies.

Dr. Sharon O'Brien is a lecturer in translation and language technology in the School of Applied Language and Intercultural Studies, Dublin City University. Her research focuses on the interaction between translators and technology, cognitive aspects of translation, research methods, including eye tracking and keyboard logging, localisation and content authoring.