ACCEPT

SEVENTH FRAMEWORK PROGRAMME

THEME ICT-2011.4.2(a)

Language Technologies

ACCEPT

Automated Community Content Editing PorTal

www.accept-project.eu

Starting date of the project: 1 January 2012

Overall duration of the project: 36 months

Data and report from user studies – Year 2

Workpackage n° 7	Name: Monolingual Postediting
Deliverable n° 7.1.2	Name: Data and report from user studies – Year 2
Due date: 31 December 2013	Submission date: 19 December 2013
Dissemination level: PU	
Organisation name of lead contractor for this deliverable: University of Edinburgh	

The research leading to these results has received funding from the European Community's Seventh Framework Programme (*FP7/2007-2013*) under *grant agreement n° 288769.*

Contents

Data and report from user studies – Year 2

Objectives of the Deliverable

WP7 aims at exploring how monolingual speakers of the target language can be aided by machine translation systems to translate. The main objective of this deliverable is to present findings from a post-editing study of volunteer participants of the Norton Community with limited source language knowledge.

Overall effort allocated to this deliverable according to description of work: 1 PM.

Introduction

This experiment was conducted with 18 volunteer participants of the German Norton Community, who post-edited 12 tasks (one question plus solution from a thread) that had been machine translated from English into German. There were two groups of post-editors for this experiment (called later Group A and Group B), editing similar but not the same content to increase the coverage of forum material and to obtain a more representative sample. Similar tasks were selected by clustering based on characteristics, such as number of words, type-token ratio etc. The tasks were edited in the ACCEPT portal using the post-editing functionalities. The tasks were configured to monolingual post-editing by default but included the option of displaying the source. The decision to introduce a switch was based on a previous study, in which participants requested to see the source. For each new task, access to the source had to be switched on again. Participants were able to use the switch on a segment level. Of the 18 participants, "Editor1" was the only participant who chose to edit completely monolingually (i.e. who switched off the English source for all tasks). There were no participants who switched off the source for at least one or more tasks. Thus, data for monolingual post-editing was only collected for this participant. The following shows the results for the monolingual post-editor in comparison to the bilingual post-editors in the same group.

		FR-EN	EN-FR	EN-DE
Symantec	Monolingual	0	0	1
	Bilingual	0	0	17

Table 1. Experimental setup (number of participants)

Pre-Task Questionnaire Analysis

The pre-task questionnaire revealed the following. Table 2 shows that out of 18 participants, 13 (72.2%) were male and 5 (27.8%) were female. The monolingual post-editor is male, belongs to the "35-44" category and is a regular member of the German Norton Community, who has asked questions but not provided answers.

age	female	%	male	%	Total	%
18-24	0	0	1	5.56	1	5.56
25-34	2	11.1	2	11.1	4	22.2
35-44	1	5.56	6	33.3	7	38.86
45-54	1	5.56	2	11.12	3	16.66
55-64	1	5.56	1	5.56	2	11.12
over 65	0	0	1	5.56	1	5.56
Total	5	27.78	13	72.22	18	100

Table 2. Age and gender distribution

Table 3 shows the self-reported English (reading) and German (writing) skills, according to the Cedefop descriptions. The English skills were rather mixed with a tendency towards C1 and C2, the two most advanced language skill categories. The German skills were concentrated on the B and C categories. The monolingual post-editor rated his knowledge of English as A2 (second worst category) and his knowledge of German as C2 (best category).

	English Skills (reading)	%	German Skills (writing)	%
A1 (worst)	0	0%	0	0
A2	3	17%	0	0
B1	3	17%	2	11%
B2	2	11%	2	11%
C1	5	28%	3	17%
C2 (best)	5	28%	11	61%
Total	18	100.00	18	100

Table 3. English and German skills distribution

Evaluation

TER Results

	Group A (MT vs. PE)	Group A (PE vs. Ref.)
Editor 1	0.3309	0.7838
Editor 2	0.4290	0.6772
Editor 3	0.4916	0.7179
Editor 4	0.4345	0.6953
Editor 5	0.4257	0.6516
Editor 6	0.5964	0.6385
Editor 7	0.4578	0.7169
Editor 8	0.4749	0.6757
Editor 9	0.4345	0.7315
Reference	0.7632	n/a

Table 4. TER results comparing MT output with PE results and PE results with reference translation

Table 4 shows how much of the MT content was changed by the post-editors (for group A) represented in TER scores in the first column and how far the post-edited versions differ from the reference translation in the second column. The highest and lowest scores are underlined for both categories. It is evident that the reference translations have the highest TER scores and thus differ from the MT results the most, which is due to the fact that they were created without access to the

machine translated output. For group A, Editor 1 (the monolingual post-editor) is the one who changed the MT output the least and is the furthest away from the reference translation, while Editor 6 is the one who changed the MT output the most and is closest to the reference translation at the same time. There is a strong negative correlation between the number of changes and the closeness to the reference translation with a correlation coefficient of -0.7, i.e. the less is changed the more the content differs from the reference translation.

Productivity

Table 5 presents the user productivity measured in words (source text) per hour for group A. The "most" productive and the ''least" productive editors are underlined. Productivity ranged from 348 words per hour to 2451 words per hour, with an average of 1093 words per hour. In this case, the monolingual post-editor was 2.24 times as fast as the bilingual post-editor (on average), which is plausible, as a monolingual post-editor does not spend time referring to the original text.

	Group A
Editor 1	2450.74
Editor 2	1354.24
Editor 3	522.47
Editor 4	348.08
Editor 5	1756.54
Editor 6	840.65
Editor 7	1163.3
Editor 8	745.31
Editor 9	653.28
average	1092.73

Table 5. User productivity in words (ST) per hour with the least productive and the most productive participant highlighted

Linguistic Annotation

Table 6 displays the absolute number of errors and the percentages in the raw machine translation output divided into the categories "Accuracy", "Language" and "Terminology". Errors classified under the Accuracy category denote translation errors, which are normally detected by comparing the source and target texts (omission/addition, untranslated text, incorrect meaning, etc.). Errors under the Language category denote language errors. Usually, these are deviations from generally accepted language conventions (punctuation, spelling/typo, and grammar/syntax). Errors classified under the Terminology category denote errors due to a bad lexical choice. Usually, these are deviations from sector- or context-specific terminology.

	MT (Group A)	%	MT (Group B)	%
Accuracy (A)	295	**47.2**	235	38.65
Language (L)	306	**48.96**	355	58.39
Terminology (T)	24	**3.84**	18	2.96
Total	625	100	608	100

Table 6. Error distribution in raw machine translated output

Table 7 shows summarised results of the error annotation of a sample of the content for three post-editors of group A. Comparing the percentages of errors present in the raw MT output and the post-edited output, it is evident that Editor 1, the monolingual post-editor, produced or failed to correct considerably more accuracy errors and slightly more language errors than the bilingual post-editors. This may be the case because the Editor 1 left segments untouched, which also explains the closeness expressed in TER scores for him in Table 4.

	Editor 1 (group A)	Editor 2 (group A)	Editor 3 (group A)
Number of sentences	72	74	71
Number of words	864	971	859
Number of errors	165	78	86
Average errors per	2.292	1.054	1.211
Sentences with 0 errors	14	22	31
…1 error…	18	39	18
…2 errors…	12	5	8
…3 errors…	14	3	10
…4 errors…	5	5	3
…5 errors…	3	0	1
…6 errors…	2	0	0
…7 errors…	1	0	0
…8 errors…	1	0	0
…9 errors…	1	0	0
…10 errors…	1	0	0
Accuracy (A)	87 (**53%**)	15 (**19.23%**)	30 (**34.89%**)
Language (L)	76 (**46.06%**)	56 (**71.79%**)	47 (**54.65%**)
Terminology (T)	2 (**1.21%**)	7 (**8.98%**)	9 (**10.47%**)

Table 7. Error annotation results (sample)

Conclusion

From the data that was collected in this experiment, it can concluded that this particular set-up – English-German as a language pair, Norton Community, option to edit bilingually/monolingually, volunteer post-editors – does not support the concept of monolingual post-editing well. Especially the (poor) quality of the raw MT output seems to negatively influence the results and user motivation. To explore other directions for monolingual post-editing, a monolingual study in the English Norton Community with French-English as a language pair is envisaged for Y3 (Symantec). This involves the integration of a paraphrase functionality developed by UEDIN (to provide the editors with a larger choice of phrases to choose from in order to deal with mistranslation easier and with greater certainty, cf. DOW Task 7.2) and the use of Acrolinx post-editing rules (Task 2.2). It is expected that the raw MT output will be of better quality, due to the language pair, and, for the use case to be of bigger impact, as the community of the target language is of considerably larger size than the German and French Norton Communities.