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 on monolingual post-editing, and report on comparison of bilingual and monolingual editing metrics

Workpackage n° 7 & 8	Name: Monolingual Postediting, Bilingual Postediting
Deliverable n° 7.1.3 & 8.2	Name: Data and report from user studies on monolingual post-editing, and report on comparison of bilingual and monolingual editing metrics
Due date: 31 December 2014	Submission date: 19 December 2014

Dissemination level: PU

Organisation name of lead contractor for this deliverable: University of Edinburgh, Lexcelera

Author(s): Ulrich Germann, Linda Mitchell, Laurence Roguet

Internal reviewer(s): Pierrette Bouillon, Victoria Porro Rodriguez

Proofreading: Manny Rayner, Violeta Seretan

Copyediting: Violeta Seretan, Barry Haddow

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.*

SEVENTH FRAMEWORK PROGRAMME

Contents

Data and Report from User Studies on Monolingual Post-editing, and Report on Comparison of Bilingual and Monolingual Editing Metrics

1 Overview

This report jointly covers Deliverables D7.1.3: *Data and Report from User Studies on Monolingual Post-editing – Year 3* and D8.2: *Report on Comparison of Bilingual and Monolingual Editing Metrics*, as laid out in the ACCEPT Description of Work. The merge of these deliverables was agreed to by the Project Officer as of 4 November 2014.

The goal of WP 7: *Monolingual Post-editing* was to evaluate the feasibility of monolingual post-editing, where post-editors are expected to be competent in the target language but not necessarily in the source language, and have no access to the source of the translation. In the Year 2 user studies on monolingual post-editing, we found that post-editors who have at least some knowledge of the source language very strongly prefer to have access to the source text. Because of this, and our conjecture that raw MT quality is a strong determining factor in the feasibility of monolingual post-editing, we decided to complement the user study on monolingual post-editing of Symantec Forum data specified in the Description of Work with a study on monolingual post-editing of longer texts (translated from French into English) in the medical / humanitarian relief and field work domain, using post-editors who are native or near-native speakers of English, and who are not (necessarily) bilingual. The texts chosen for this study were identical to those used in the corresponding bilingual study on medical text, which is covered by Deliverable D8.1.3. The choice of languages for this study was driven by the availability of volunteer translators for the bilingual counterpart of the study and the fact that machine translation from French into English generally performs reasonably well.

2 Symantec Monolingual and Bilingual Post-Editing User Study

2.1 Introduction

This experiment aimed at exploring the impact of source language access on the post-editing process when post-editors are provided with editing assistance technology including target text checking and paraphrasing. With this experiment we sought to establish any differences in quality and behaviour across the four following setups: EN>FR (access to source, no access to source); FR>EN (access to source, no access to source). Section 2.2 describes how the post-editing experiments were designed. Section 2.3 presents the analysis of the post-edited data, while Section 2.4 presents the data collected in the post-task survey. Section 2.5 outlines the evaluation results of the post-edited data consisting of TER results and the human evaluation results.

2.2 Experiment Design

The Symantec experiment was conducted with 13 volunteer participants, some members of the English and French Norton Communities, and some Symantec employees, who post-edited 5 tasks (30 segments). Each task consisted of a forum post (including its subject line). The posts selected for the tasks originated from the "Baudolino" study described in Deliverable D6.5. The language pairs under scrutiny were English-French and French-English. There were four groups of post-editors for this experiment, with either three or four post-editors per setup (monolingual: access to source/bilingual: no access to source) and language direction (English-French/French-English).

The four setups are presented below:

1. Monolingual, French-English:
   (a) 459 words (30 segments)
   (b) 4 post-editors (EN1 - EN4)
2. Bilingual, French-English:
   (a) 459 words (30 segments)
   (b) 3 post-editors (FR-EN1 - FR-EN3)
3. Monolingual, English-French:
   (a) 349 words (30 segments)
   (b) 3 post-editors (FR1 - FR3)
4. Bilingual, English-French:
   (a) 349 words (30 segments)
   (b) 3 post-editors (EN-FR1 - EN-FR3)

The post-editing tasks took place in the ACCEPT portal, which contained projects configured for users to be able to access the following form of assistance: a "Symantec" paraphrasing service provided by UEDIN based on the approach described in Deliverable D7.2 and an interactive checking service (grammar and spelling) provided by Acrolinx, using the rule set Postediting-EN-FR for the EN>FR language pair and the rule set Postediting-FR-EN for the FR>EN language pair, as described in Deliverable D2.4.

2.3 Analysis of Post-Edited Data

This section focusses on an empirical overview of the post-edited data collected, as extracted from the XLIFF post-editing reports.[1] This includes the post-editing times, the frequency of assistance used, i.e. paraphrasing and interactive checking, and the average number of key strokes per setup.

[1]The data is available at `http://www.accept.unige.ch/Products/D7-1-3-data-uedin.zip`.

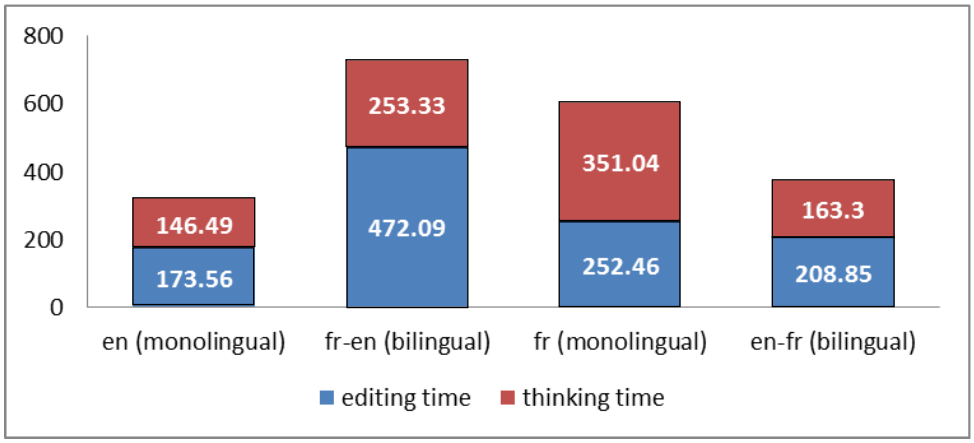

Figure 1: Average time (total, editing, thinking) per task and language direction

Figure 1 shows the average time for the two setups and the two language directions that the post-editor respectively spent editing or typing (blue – lower) and idle, hence presumably thinking (red – upper).

There seems to be no clear pattern as to whether monolingual or bilingual post-editing took more time. It seems to be generally the case that the editing time is longer than the thinking time.

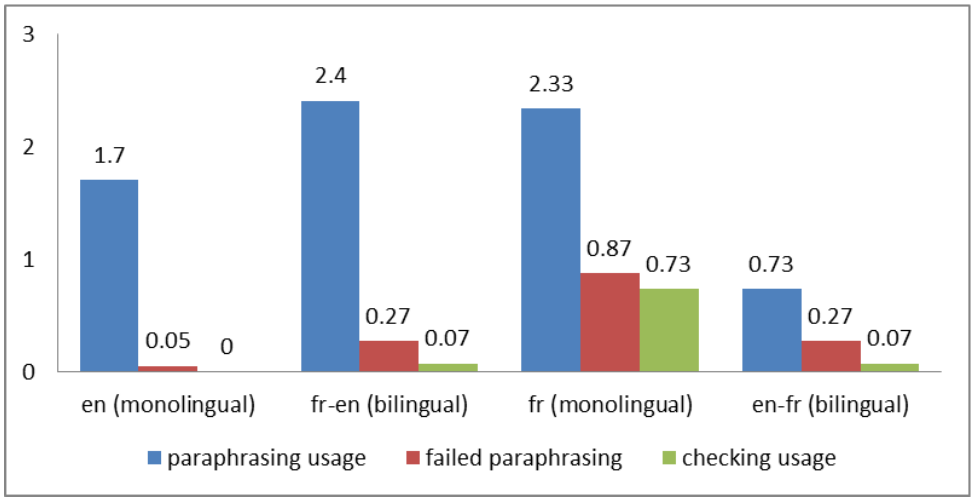

Figure 2: Average number of form of assistance used per task and language direction

Figure 2 displays the average number of times assistance was used per task (approx. 6 segments), consisting of paraphrasing used and paraphrasing failed, as well as the number of times the ACCEPT interactive checking was triggered. Paraphrasing was on average triggered between one and two times per task (blue bar – left). Thus, paraphrasing was requested in 16% to 33% of the segments. Of these, an average of much less than one paraphrasing attempt per task failed (red bar – middle). The ACCEPT interactive checking was on average used a little less than once per task for the French monolingual setup, and hardly at all for the other setups (green bar – right).

Post-editor	Paraphrasing requests	Paraphrasing failure	Interactive checking
EN1	21	0	0
EN2	2	0	0
EN3	9	1	0
EN4	2	0	0
FR-EN1	15	0	1
FR-EN2	9	1	0
FR-EN3	12	3	0
FR1	10	4	2
FR2	18	7	8
FR3	7	2	1
EN-FR1	8	3	0
EN-FR2	0	0	0
EN-FR3	3	1	1

Table 1: Total number of assistance triggered (paraphrasing, checking) per post-editor

Table 1 shows the 13 users and their total use of each form of assistance (paraphrasing and checking) for their setups. The total number of paraphrasing requests triggered per post-editor ranges from 0 to 21 (per 30 segments) with a median of 9. Paraphrasing failure[2] ranges from 0 to 7 with a median of 1. Interactive checking ranges from 0 to 8 with a median of 0. Eight suggestions that were returned by the interactive checking service were accepted, all of which were triggered by the same user in the Monolingual EN-FR setup. This involved the following changes: punctuation (2 instances), spacing (2 instances), deleting a redundant word (3 instances), capitalisation (1 instance). The paraphrasing feature was frequently used in both monolingual setups, as well as for the bilingual setup FR-EN. However, for EN>FR, it would seem that monolingual users triggered the paraphrasing assistance more frequently than their bilingual counterparts. As pointed out above, this is not the case for the other language direction. It can be concluded that the form of assistance used seems to only somewhat depend on factors of the experimental setup (language direction, access to source etc). It should be noted that the second column in Table 1 shows all paraphrasing requests. Of these, only 7 were accepted by the post-editors in total (3 for English and 4 for French).

Figure 3 shows the average number of keystrokes per task and setup. It is evident that there is no large difference between the average total number of keystrokes across all setups. Monolingual post-editors produced output with slightly fewer keystrokes than bilingual post-editors. The largest number of keystrokes occurred for the bilingual setup with English as a target language.

[2]Paraphrasing failure includes all instances for which the UEDIN paraphrase server returns an error or when it does not respond within a pre-defined timeframe.

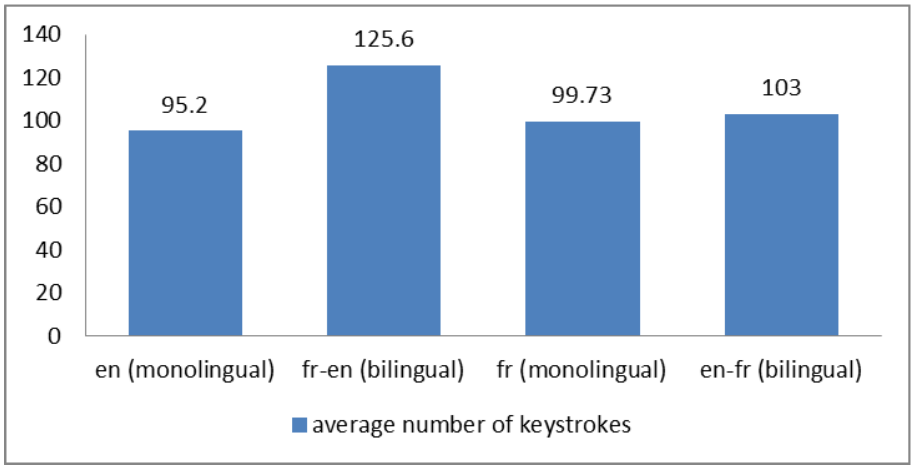

Figure 3: Average number of keystrokes per task and language direction

2.4 Survey Data

The following section summarises data elicited in a post-task survey, focussing on the ACCEPT post-editing environment:

a. Which feature did you like best in the editor? Why?
b. Which feature did you like least in the editor? Why?
c. What could be improved in the editor?

Table 2 displays the answers to the three questions presented in summarised form.

a. best	#	%	**b. least**	#	%	**c. improved**	#	%
none	3	23	none	6	46	no answer	2	15
easy to use	3	23	suggestions by highlighting	3	23	change highlighting	2	15
access to source text	1	8	lack of formatting	1	8	collaborative editing	1	8
navigation	2	15	automatic correction	1	8	contextual translation	1	8
ability to comment	1	8	underlining	1	8	more features	1	8
ability to change content between paragraphs	1	8				improve dismissing tips	1	8
manual translation	1	8				add user formatting	1	8
						return results for long sentences	1	8
						improve translations	1	8
						user interface	1	8

Table 2: Post-editor attitudes on post-editing environment

It is evident that the post-editors were happy with or indifferent towards the editor for the most part. The ease of use was emphasized on the positive side, as was the highlighting as triggering the paraphrasing suggestions on the negative side. There were various suggestions on how to improve the editor, which were largely based on rectifying the issues mentioned under b.

2.5 Evaluation Results

The following section discusses the quality of the post-edited output by focussing on the TER results, firstly, and the results from a human evaluation step, secondly.

2.5.1 TER Results

This section presents the evaluation results. The first metric used is TER, which provides a quantitative measure of how much on text has changed based on another text. Here, it measures the difference between the post-edited text and the MT output. A high value of TER indicates a large difference and a small value a low difference. With this comparison, we obtain an estimate of the amount of work performed by the post-editors. Table 3 presents TER results, showing PE versus MT.

Setup		**1**	**2**	**3**	**4**	**Average**
bilingual	EN-FR	0.12	0.19	0.27	n/a	0.19
monolingual	FR	0.08	0.07	0.31	n/a	0.15
bilingual	FR-EN	0.07	0.30	0.23	n/a	0.20
monolingual	EN	0.10	0.14	0.18	0.08	0.13

Table 3: TER results per setup and language direction comparing PE to MT output

It is evident that post-editors implemented slightly more changes for both bilingual setups, compared to the monolingual ones. A reason for the smaller number of changes implemented in the monolingual setups could be that the post-editors were unable to correct all errors, due to lack of access to the source text. In general, it is striking that relatively little editing was performed in all groups.

2.5.2 Human Evaluation Results

The evaluation of the MT segments and post-edited segments took place in the ACCEPT Portal using the internal evaluation project functionality described in Deliverable D5.8. The following four evaluation projects were created:

- Monolingual English>French
- Bilingual English>French
- Monolingual French>English
- Bilingual French >English.

Each project contained 20 to 25 evaluation tasks, as there were 5 post-editing tasks associated with 4 or 5 users (one of these "users" was the MT system and the others were the post-editors). Each project was configured such that duplicated segments (e.g. when a target MT segment was the same as a target post-edited segment, or when two target

post-edited segments were identical) would not be rated twice by a given evaluator. This was done to ensure that the evaluation process would not be too repetitive for evaluators. The evaluation was conducted using three Norton forum users with advanced to native-level French and English skills. Each user was invited to each of the four evaluation projects and was told to complete the evaluation tasks in any particular order (the origin of the target segments being hidden). The evaluation guidelines are shown in Figure 4.

Always read the target text first (in the middle column), then read the source text in the left column. Then choose one of the four quality ratings for the target text:

- Very good: Your understanding of the target text first is not improved by the reading of the source text because the target text is satisfactory and would not need to be modified (grammatically correct/proper terminology is used/maybe not stylistically perfect but fulfils the main objective, i.e. transferring accurately all information).
- Good: Your understanding of the target text first is not improved by the reading of the source text even though the target text contains minor grammatical mistakes (e.g. short distance word order/punctuation errors/word formation).
- Bad: Your understanding of the target text first is improved by the reading of the source text, due to significant errors in the target text (e.g. long distance word order/textual and syntactic coherence/ word formation).
- Very bad: Your understanding only derives from the reading of the source text, as you could not understand the target text. It contained serious errors in any of the categories listed above.

Figure 4: The evaluation guidelines

For the analysis of the results, these categories were mapped to the following values: 4, 3, 2, 1. The following number of ratings was obtained, showing a larger amount of repetitions in the English>French monolingual project (possibly due to the fact that post-editors were not sure how to edit the target text without access to the source text):

- Monolingual English>French: 234
- Bilingual English>French: 246
- Monolingual French>English: 252
- Bilingual French>English: 255.

In order to calculate final average results for each user (either MT or post-editor), repeated segments were assigned the score given to one of the repetitions (to avoid any score imbalance across users). Table 4 shows the final average scores ordered per average score per target language.

Language pair	Post-editor	Average score	Project type
English>French	EN-FR3	3.73	Bilingual
English>French	EN-FR1	3.52	Bilingual
English>French	EN-FR2	3.51	Bilingual
English>French	FR1	3.19	Monolingual
English>French	FR2	3.15	Monolingual
English>French	FR3	3.13	Monolingual
English>French	MT	3.09	Both
French>English	FR-EN1	3.6	Bilingual
French>English	FR-EN3	3.56	Bilingual
French>English	EN2	3.52	Monolingual
French>English	EN3	3.38	Monolingual
French>English	FR-EN2	3.17	Bilingual
French>English	MT	3.13	Both
French>English	EN1	3.1	Monolingual
French>English	EN4	3.04	Monolingual

Table 4: Human evaluation results

2.5.3 Analysis

For both language pairs, the content produced by bilingual post-editors received on average higher quality scores than the content produced by monolingual post-editors. The quality of the post-edited content seemed to greatly benefit from access to the source (bilingual post-editing). This is evident for the FR>EN language pair, for which two of the monolingual post-editors produced content of slightly inferior quality (on average) to that of the initial MT output. However, it is worth highlighting the work produced by EN2, since the quality resulting from this work (performed in a monolingual context) is almost as high as the one produced by FR-EN1 and FR-EN3 in a bilingual context (3.52 vs 3.6/3.56).

This confirms the usefulness of monolingual post-editing as long as it is performs by a post-editor with sufficient domain and language skills. Furthermore, there was no correlation between the TER scores as presented in Table 3 and the quality scores from the human evaluation as presented in Table 4.

In general, the average score for post-edited content did not reach more than 3.73 points (out of 4), which shows that, while post-editing can dramatically improve the quality MT output, there are still a few segments the quality of which could be further improved (in either language direction and setup). This suggests that sequential post-editing could be envisaged so that post-edited segments of imperfect quality can be further improved by other post-editors.

3 Post-editing in the Medical Domain

The TSF experiment was conducted with 16 volunteer participants from UEDIN using the task of monolingual post-editing into English. For bilingual post-editing there were 12 volunteer participants from the University of Rennes (CFTTR) and 4 volunteer participants from the TSF community. These two groups covered the language pairs English-French and French-English. It should be noted that, although the French-speaking group was originally split into subgroups respectively assigned monolingual and bilingual post-editing, all the subkects assigned monolingual tasks switched to bilingual ones. Two sets of source-language content were used, consisting of around 5,000 words per set divided into 10 tasks of 500 words each. The same content was used for monolingual and bilingual post-editing.

3.1 Setup

The study on monolingual post-editing in the medical domain was conducted at the University of Edinburgh. We recruited 16 paid post-editors, most of them students at the University of Edinburgh. Payment was on an hourly basis, with a generous limit of 10 hours in total to complete the tasks (4241 words of MT output). Recruitment criteria for participants were:

- native or near-native competence in writing English;
- familiarity with medical terminology (although it turned out that there was hardly any medical terminology in the texts so specific that it would have required medical background knowledge or even medical training).

The test material was taken from two documents, one a report by *Doctors without Borders* on emergency relief efforts in Haiti after the outbreak of cholera, the other from an instructional manual on baby massage. The material was split into ten segments ("tasks") of ca. 350–600 words each. The first task was intended as a warm-up, to familiarise the participants with the ACCEPT tool, and not considered for evaluation. Participants were briefed about the task and given an introduction to the ACCEPT tool in a short introductory session on-site; afterwards they completed the task on-line at times and locations of their preference and choice. All tasks were completed within a window of one to two weeks. A comment box under the text edit window in the ACCEPT tool allowed participants to report on their work as they were progressing. We encouraged them to use it to keep track of when they resorted to external resources (such as Google, on-line thesauri etc.) while editing, and to freely comment on the post-editing process.

This study contrasted two conditions:

- monolingual post-editing without further assistance to post-editors;
- monolingual post-editing with paraphrase assistance.

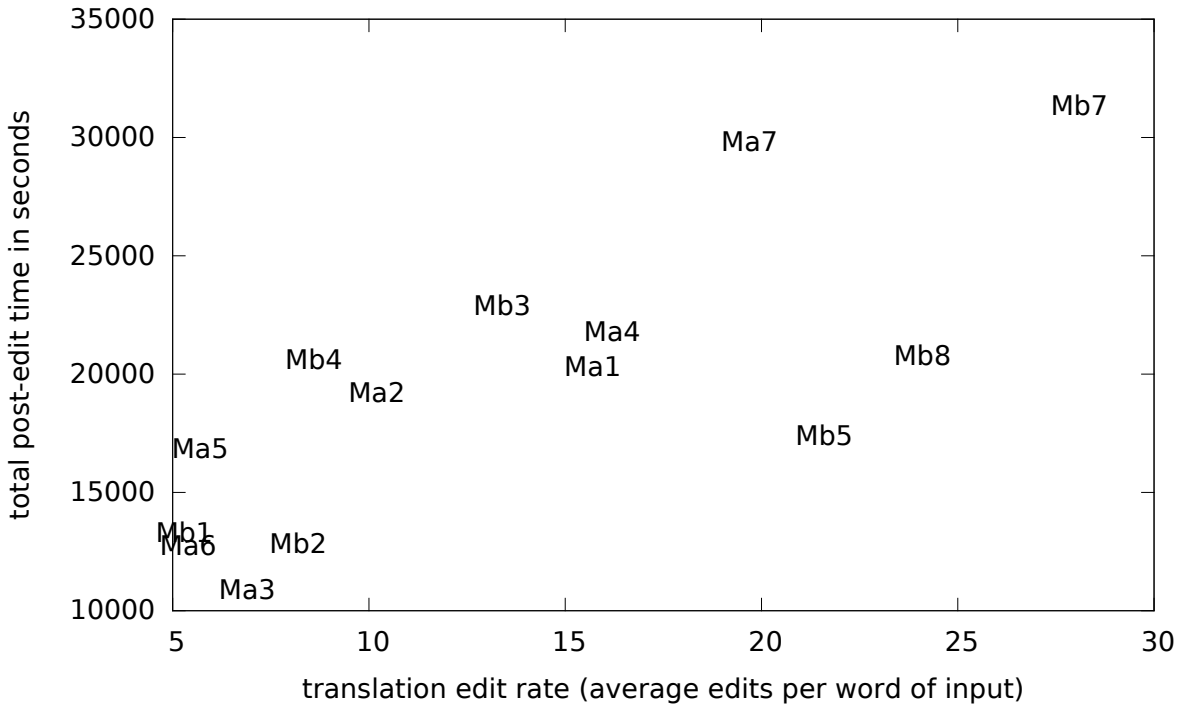

Figure 5: Translation edit rate (x-axis, in %) vs. total post-edit times (y-axis, in seconds) for 14 participants in the monolingual post-editing study in the medical domain.

Machine translation was produced from input pre-edited with pre-editing rules developed under Work Package 2 (cf. Deliverable D2.2). The paraphrasing assistance is described in Deliverable D7.2. In a nutshell, when the paraphrasing feature is activated, highlighting a short text passage triggers a lookup request to a paraphrase server that provides translation alternatives by re-translating the passage using a dictionary of paraphrases derived from the original training data while considering the local context of the highlighted phrase in question. The user is then presented with a short list of the top-ranked alternatives.

The 16 post-editors were split into two groups. Post-editors in condition A had access to the paraphrasing tool in the first 5 of the 9 tasks that the actual study comprises, and no support in the last four; for participants in Group B conditions were flipped.

3.2 Analysis of Timing Data

Figure 5 shows a scatter plot of total post-edit times vs. translation edit rate for 14 of the participants in the monolingual post-editing experiment in the medical domain.[3] The plot suggests that there are two types of editors: "fast" editors who change little (Ma3,5,6;Mb1,2), and "slow" editors who tend to edit more and spend more time (Ma1,4,7;Mb5,7,80). Ma2 and Mb4 are borderline cases. Within each group, the correlation between edit time and edit rate is low, and both edit time and edit rate show considerable variance across post-editors.

[3]Two participants completed only about half of the tasks and were omitted from consideration here.

In order to understand where post-editors spend their time, we plotted the total edit time per segment for each of the 14 participants who completed all Tasks in the project. The plots are shown in Figures 6 (Group A) and 7 (Group B).

The x-axis represents the segments of the nine tasks in linear sequence, the y-axis measures the number of seconds *per word* spent on the particular segment. Dotted lines indicated task boundaries; the dashed line marks the topic boundary between "cholera relief efforts" and "baby massage".

The plots allow a number of observations. First, notice the high volatility of time spent on each segment. The total time spend on the post-editing task is clearly dominated by relatively few segments that the post-editors spent a lot of time on. It is not clear what exactly triggered these pauses. There may have been external interruptions (e.g. phone calls, short breaks from work). Judging from comments that participants left in the tool's comment box, a good number of these long "pauses" can, however, also be attributed to on-line searches for clarification. For example, several participants reported that they googled the term "Plumpy'Nut" (which, by the way, is the brand-name of "a peanut-based paste in a plastic wrapper for treatment of severe acute malnutrition" (Wikipedia)). Especially the text on cholera relief efforts also contained a plethora of acronyms that post-editors chose to research and verify on-line.

Second, we do not observe any speed-up over time. One might expect that, as users become more familiar with the tool, their productivity would increase. However, the logged data do not support that conjecture. Our interpretation of this observation is that issues relating to text understanding (independent of the interface) are the dominating effect. That said, we should mention that post-editors did not like the lack of formatting in the whole-text panel on the left of the interface (from which the user selects sentences to edit). Several of them raised this as an issue that impeded their work, especially when it came to recognising section titles, bullet points, etc., as such.

Third, there is no clear correlation between post-editors as to when the observed delays occur. This may be partly due to the way timing information is captured and assigned to the segment currently in the edit window, irrespective of where the post-editors attention is — they might be reading the full text in the left column of the tool. Notice that some but not all post-editors seem to have a revision phase at the end of some tasks. Post-editors Ma2, Ma5, Ma6, for example, show some long delays just before task boundaries.

3.3 Use of Paraphrasing Support

The top parts in Tables 5 and 6 summarize the use of paraphrasing assistance by the post-editors. The frequency of use varies widely. Ignoring Post-editors Ma8 and Mb6, who happened not to edit the set of tasks in their respective conditions that provided paraphrase support, most of the participants asked for paraphrases a few dozen times (a good number of requests certainly submitted by way of trying out the assistance than rather being in need of it at the time), but had little success in finding useful suggestions. A few participants, however, used the tool more frequently with somewhat higher but overall rather low success rate. The most frequent use was by Post-editor Mb1 (167 lookup

post-editor	Ma1	Ma2	Ma3	Ma4	Ma5	Ma6	Ma7	Ma8
paraphrase support:								
total user requests	17	39	2	99	55	27	44	0
no server response	1	2	1	16	12	8	5	0
no paraphrases found	6	5	0	37	7	16	26	0
no suggestion chosen	8	32	1	36	29	2	11	0
success	2	0	0	10	7	1	2	0
source words	4496	4496	4496	4496	4496	4496	4496	2206
MT output (wrds)	4241	4241	4241	4241	4241	4241	4241	2115
final version (wrds)	4234	4320	4210	4184	4229	4273	4232	2162
total segments	266	266	266	266	266	266	266	170
segm. with no edits	98	125	153	100	174	195	96	64
total # of edits	673	432	294	687	243	229	837	321
TER	15.7%	10.2%	6.9%	16.2%	5.7%	5.4%	19.7%	15.2%
total thinking time	3:32:36	4:00:05	2:02:43	3:51:52	3:07:55	2:45:28	5:52:30	1:53:20
total typing time	2:05:17	1:19:30	0:58:40	2:11:08	1:32:27	0:46:46	2:24:02	1:33:22
total post-edit time	5:37:54	5:19:36	3:01:24	6:03:01	4:40:22	3:32:15	8:16:32	3:26:42

Table 5: Post-editing statistics in Condition A, with paraphrase support offered for the first five of the tasks. Post-editor Ma8 completed only the last four tasks, which didn't offer any support. Note that the TER reported in this table is a measure of how much the text was edited, not a measure of final text quality.

requests, with a success rate of ca. 13%). The most successful use of the tool was by post-editor Mb3 with a success rate of 23% (22/92). Overall, most the study participants did not consider the paraphrasing tool particularly useful. In a post-study questionnaire, only two participants rated the paraphrasing support as "somewhat useful" in that it helped them find better translations. 8 didn't see the need for paraphrase assistance, whereas 6 found the tool unhelpful because of the poor quality of the suggestions it returned. None of them found the tool "very useful", which was the fourth option offered as an answer in the multiple-choice question.

However, all participants reported the use of online resources to aid the post-editing process — a practice that we explicitly allowed to test the post-editing process under realistic conditions. All but one reported the use of Google; several also resorted to on-line resources such as on-line dictionaries and therauri to find synonyms. One thing that was particularly noted in the participants' comments was the need to understand acronyms, which permeated the text on cholera relief efforts. So while the paraphrasing engine in place did not convince most of the users in terms of quality, paraphrasing / synonym support per se can still be considered a useful form of assistance to post-editors.

post-editor	Mb1	Mb2	Mb3	Mb4	Mb5	Mb6	Mb7	Mb8
paraphrase support:								
total user requests	167	22	92	119	65	0	15	40
no server response	0	0	0	0	8	0	4	0
no paraphrases found	37	8	7	25	33	0	7	15
no suggestion chosen	108	13	64	80	22	0	4	22
success	22	1	21	14	2	0	0	3
source words	4496	4496	4496	4496	4496	2290	4496	4496
mt output (wrds)	4241	4241	4241	4241	4241	2126	4241	4241
final version (wrds)	4220	4223	4224	4266	4179	2134	4221	4305
total segments	266	266	266	266	266	96	266	266
segm. with no edits	167	121	90	153	91	58	27	104
total # of edit op's	226	349	570	366	918	113	1193	1024
TER	5.3%	8.2%	13.4%	8.6%	21.6%	5.3%	28.1%	24.1%
total thinking time	1:55:45	2:29:11	4:33:59	3:29:00	2:03:34	0:24:48	4:24:49	1:34:04
total typing time	1:45:11	1:04:18	1:47:08	2:14:04	2:46:09	0:13:13	4:16:51	4:11:53
total post-edit time	3:40:56	3:33:29	6:21:08	5:43:05	4:49:44	0:38:01	8:41:40	5:45:57

Table 6: Post-editing statistics in Condition B, with paraphrase support offered for the last four of the tasks. Post-editor Mb completed only the last 4 tasks, which didn't offer any support. As in Table 5, the TER reported here is a measure of how much the text was edited, not a measure of final text quality. Like in Group A, one of the post-editors completed only one set of tasks instead of both.

3.4 Post-editing Effort

Post-editing effort both in terms of time spent on the task and in terms of translation edit rate varied greatly. Translation edit rates ranged from 5.3% (Mb1) to 28.1% (Mb7), without any clear correlation between time spent and Translation Edit rate. For example, Post-editor Mb1 executed only 5.3 edit operations per 100 words of "source" text (MT output) on average, spending 3 hours and 40 minutes on the entire set of tasks, whereas Post-editor Mb5 performed 4 times as many changes but spend only about 30% more time. Post-editors Mb4 and Mb8 spent about the same amount of time in total, but Mb8 made almost 3 times as many changes as Mb4. The difference in the amount of editing performed can be in part explained by different interpretation of what it means to "change as little as possible and as much as necessary to produce a fluent, natural text".

3.5 Human Evaluation

For the translations produced in the medical domain, we had human judges evaluate 105 randomly selected pairs of segments from the pools of monolingually and bilingually post-edited test data and evaluated them against a human reference translation. Of the 105 sentence pairs, 31 were considered to be of approximately equal quality by the human judges. In 45 cases the output of adapted MT plus monolingual post-editing was considered to produce better results than baseline MT plus bilingual post-editing. In 29 cases the translation was worse.

It should be noted, however, that the underlying MT systems used differed between the monolingual and the bilingual experiments. For the bilingual experiments we used the baseline systems (Deliverable D4.1). However for the monolingual experiments we used an improved version which included better tokenisation, more training data, domain adaptation through provenance features, the operation sequence model, and an improved reordering model.

4 Conclusion

The results of the Year 3 user studies can be summarized as follows.

- For language pairs where current MT technology tends to perform well in general, monolingual post-editing of automatic translations can lead to results that are on par with bilingual post-editing, according to human evaluations. However, this works only when the raw MT quality is good to begin with. It is therefore advisable to maintain topic- and domain-specific translation engines that are tailored to certain types of tasks.
- Text formatting is crucial to text understanding. Loss of text formatting can make post-editing much more difficult that it already is. This is particularly an issue as text complexity increases (tables, headlines, bullet points).
- Paraphrasing support is beneficial sometimes, but not often enough to have a clearly observable effect on post-editing effort or overall post-editing times.
- In general, there is considerable variance in post-editor behaviour, in terms of edit rates and time spent on post-editing. Moreover, the correlation between the two is poor, as for example the tables in Section 3.3 illustrate. This result is in line with the findings of other users studies on post-editing (e.g., within the EU-funded projects CasMaCat and MateCat). This makes it very difficult to draw clear conclusions with respect to the efficacy of particular post-editing tools or features that they offer.

To sum up the overall experience with exposing the tools we have developed over the course of the past three years to real-world use, we can say the following. In terms of user interface design, the tool was generally well accepted (cf. also Deliverable D8.1.3), with the caveat that it does not seem to be well suited for processing long, complex document where text formatting is important for text understanding. As far as the post-editing process itself is concerned, results are mixed. The tool works well for some people but not for all; consequently, if should be included in the toolbox of translation tools that are on offer to the translator (or post-editor), but it should not be forced upon them.

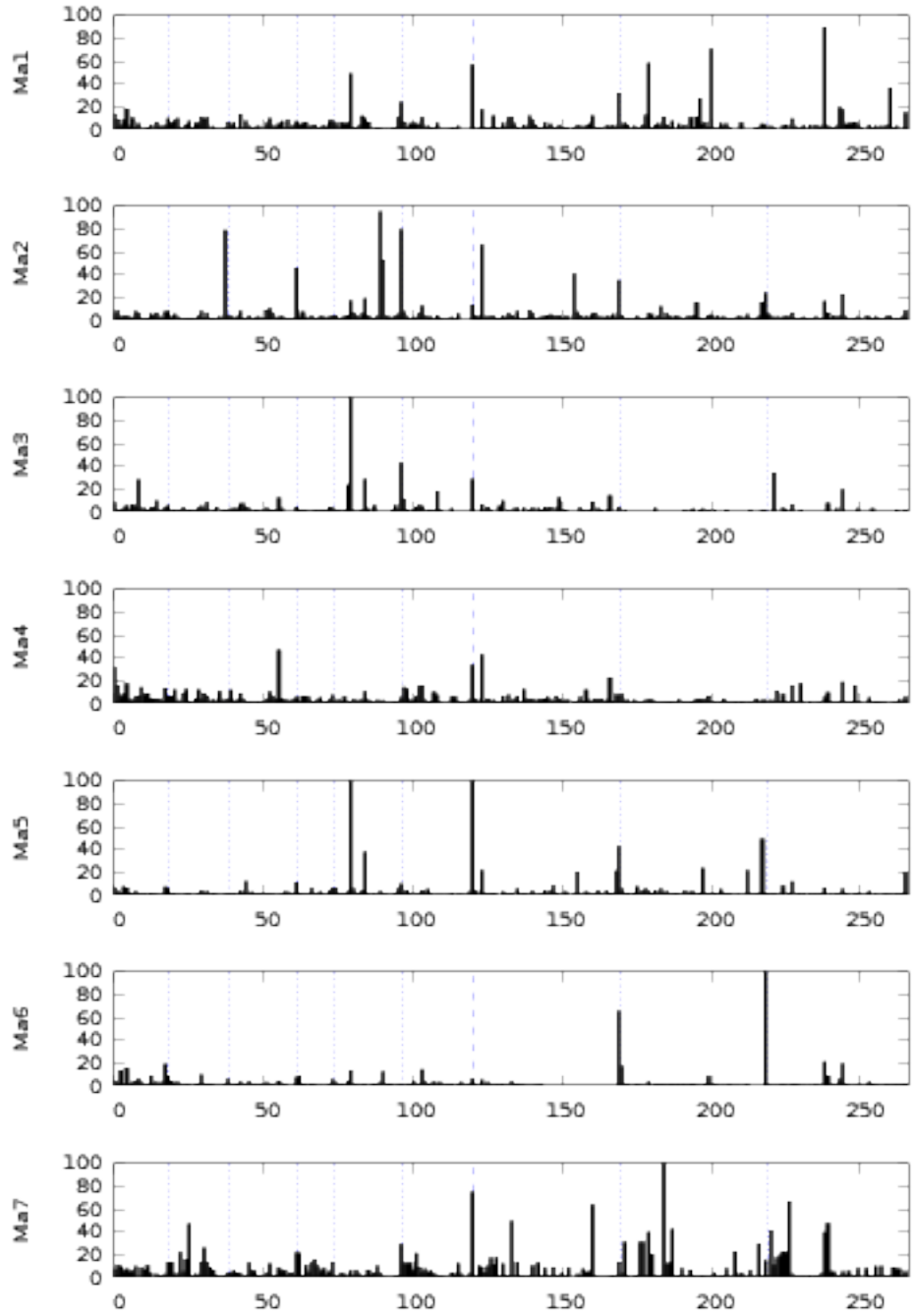

Figure 6: Total time spent by seven post-editors in Group A on editing each segment, as captured by the ACCEPT tool ("typing time" plus "thinking time"), in seconds per word of "source" text (i.e. in the case of monolingual post-editing, per word of raw MT output). The eighth post-editor completed only one of the two batches; their data are not considered in this figure.

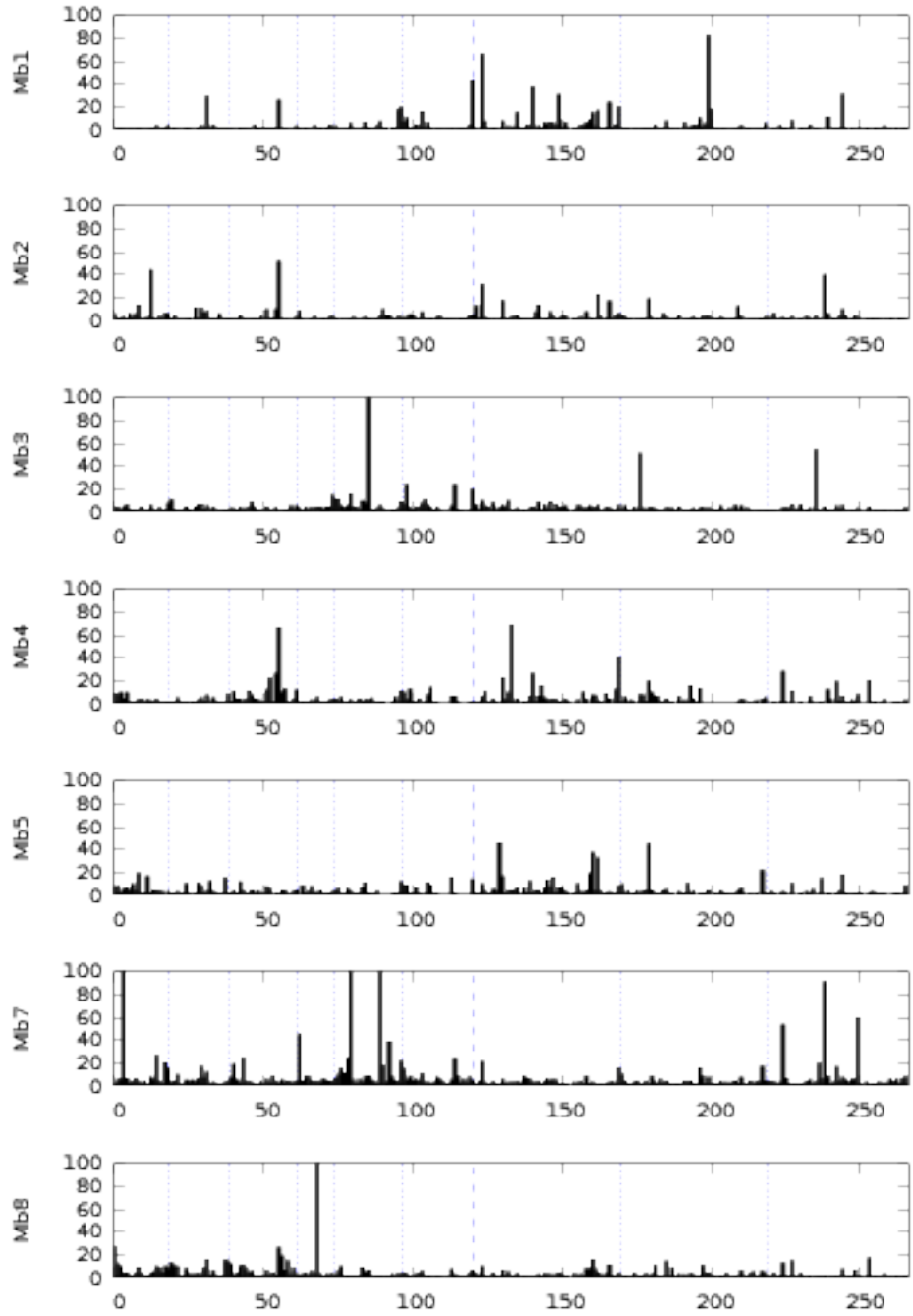

Figure 7: Total time spent by seven post-editors in Group B on editing each segment, as captured by the ACCEPT tool ("typing time" plus "thinking time"), in seconds per word of "source" text (i.e. in the case of monolingual post-editing, per word of raw MT output). The eighth post-editor completed only one of the two batches; their data are not considered in this figure.