



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 1

Workpackage n° 8	Name: Bilingual Postediting
Deliverable n° 8.1.1	Name: Data and report from user studies – Year 1
Due date: 31 December 2012	Submission date: 21 December 2012
Dissemination level: PU	
Organisation name of lead contractor for this deliverable: Lexcelera	

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*.



Content

Objectives of the Deliverable	3
Methodology for Bilingual Post-editing	3
Pre-task Questionnaire Results for TWB Community	4
Results of the Pre-task Questionnaire	4
Summary	6
Feedback during Post-editing for TWB Community.....	6
User Inputs	6
Conclusion	7
Post-task Questionnaire for TWB Community	8
Results of the Post-task Questionnaire	8
General Evaluation	8
Motivation to Edit	9
Work Environment Experience	11
Summary	11
Evaluation	12
Symantec Community	12
Manual Evaluation	12
Automatic Metrics.....	13
TWB Community	14
Manual Evaluation	14
Automatic Metrics.....	17
Timeline	18

Data and report from user studies - Year 1

Objectives of the Deliverable

The objective of workpackage WP8 is to collect user edits from users who can perform bilingual post-editing in a volunteer or community collaboration context.

This deliverable provides the first revision on the expected requirements and specifications (functional and operational, the associated plan and timetable, and the corresponding evaluation criteria, protocols and metrics).

In this deliverable, you will find all details concerning the first experiment in bilingual post-editing. The methodology that we have applied for bilingual post-editing in the Symantec and TWB community is described in the first part of this deliverable.

Methodology for Bilingual Post-editing

This first study of this work package was done in parallel by Symantec and TWB with their own community. The Machine Translation systems used in this study are the baseline MT systems developed as part of WP4.

Specificities for Symantec:

The forum posts to be post-edited were translated automatically using a local model of the MOSES translation system developed as part of WP4. The MT systems were developed with data from Symantec (English>French, English>German). For both language pairs (EN-DE, EN-FR), 4 native speakers of the target language were recruited. Users were recruited within the forums by the forum administrators. The targeted users were native speakers of the target language with enough knowledge of the source language (English) to be able to perform bilingual post-editing tasks.

For French, 28 'gurus' or active users were invited via private message. Out of these, 17 did not reply, 5 users declined and 6 accepted. One person could not be included because the targeted user profile was not met; the results for another user had to be discarded due to technical challenges. For German, 12 gurus/active users were contacted directly. 2 users did not reply and 2 users declined; 8 users agreed to take part in the study, out of which 4 had to be discarded because they did not fit the profile (e.g. their mother language was not German). The final success rate for the initial recruitment round was thus 33% for German and 15% for French, with, 4 users for German and 5 users for French.

The texts for post-editing were taken from the English-speaking support forum. They consisted of the original question in a thread and its subject line, followed by the post that had been marked as the solution to the question in the forum. The content to be post-edited was taken from a set of 347 posts, which had been extracted previously for the purpose of machine translation. There was a human translation available for use as a reference.

Specificities for TWB:

The TWB documents to be post-edited were translated automatically using a local model of the MOSES translation system developed as part of WP4. The MT system was developed with data from TWB (English>French only).

The users for this project were 20 members from the TWB community.

This community of volunteers was built through LinkedIn. The content to be post-edited was a part of a medical manual from AMREF about diseases. This manual has been divided into 20 tasks of about 500 words.

Each member of each community post-edited translations bilingually using the post-editing environment developed in WP4. The post-editing was followed by an evaluation phase.

A post-task questionnaire was used to identify and understand:

- the user sentiment,
- their perception of the task,
- the process they applied,
- problems they encountered when carrying out the task of post-editing.





Their productivity was evaluated using information automatically collected during post-editing (number of edits, time spent on post-editing, etc.). Quality was measured with both manual evaluation and automatic metrics (METEOR, TER).

The manual had been previously translated using human translators. This translation was used as a reference to evaluate and score the MT output as well as the post-edited content.

Pre-task Questionnaire Results for TWB Community

Results of the Pre-task Questionnaire





How old are you?

under 18		0.0%
18-24		0.0%
25-30		16.7%
31-40		22.2%
41-50		33.3%
over 50		27.8%
I don't wish to specify my age.		0.0%




What country do you live in?

Netherlands, United States, England, Greece, Canada, Spain, Cambodia, Argentina, Slovenia, Ireland, France, Germany





What is your knowledge level for the healthcare field?

No knowledge.		0.0%
Basic.		38.9%
Average.		22.2%
Good.		22.2%
Fluent.		16.7%





What best describes your employment status?

full time student		0.0%
part time student		0.0%
full time job		77.8%
part time job		16.7%
retired		0.0%
I don't wish to specify		5.6%

For how long have you been a member of the community?

less than one year		44.4%
1-2 years		5.6%
3-4 years		22.2%
I don't wish to specify		27.8%

On average, how many pro bono translations per month do you usually handle?

less than 1 per month		44.4%
1-5		38.9%
6-10		11.1%
more than 10		5.6%
I don't wish to specify		0.0%

Summary

Volunteers for the first part of the experiment are active members of the community (more than 50% have been members of TWB community for more than 1 year, and handle more than 5 translations per month). They come from different countries. As active members of the TWB community and professional translators, their expectations are high.

Feedback during Post-editing for TWB Community

User Inputs

During the post-editing tasks, spontaneous feedback from post-editors was collected by the community manager:

User 1:

- "The platform doesn't use French punctuation (spaces before and after semi-colons, colons etc.), which multiplies the number of necessary corrections."
- "I have worked for a long time with the platform Lingotek, which combines the use of a general translation memory and a memory specific to the project, which results in an efficient pre-translation. With this platform, I could reach up to 10 000 words a day, because the automatic pre-translation was this good. Use of the platform is free for freelancers and NGOs."

User 2:

- "It would be useful to be able to change the size of the windows of the source sentences, the sentences being worked on and the commentaries."
 - "Downloading the source document would allow a better overview, being able to see the tables, diagrams and illustrations, paragraphs. That would make the translation/edit more homogenous and true to the original document. The best thing would be to have access to the source for the entirety of the project as well as for each of the tasks."
 - "Check-out/ check in: I can't see the formal way of accepting or reserving a task. Does clicking on 'Editor' in the table of tasks mean that the person in charge of the task in question sees it? It would be interesting to be able to visualize the content of a task before deciding whether or not to accept it. And once the task is accepted and assigned to a translator it should be marked as such so that the other translators cannot change it. It is similar to the idea of Check-in/check-out of a module or source document in software development."
 - "I also tried to edit task 10 of monolingual project TWB_Chapter2_enfr and it seems to me to be an unreasonable task. The translated text is so muddled that the editing requires imagination for some sentences. It is difficult, almost impossible to understand the mistranslation. Without access to the source text I don't think it is possible to produce a viable edited text."
 - "When you edit a segment you have the possibility of stopping the editing and returning to the original text. But once the segment has been edited and you have moved onto another segment there is no way of finding the original text again. During the proofreading of all of the text I would have liked to have been able to

reread certain parts of the original to check whether or not I missed a possible interpretation.”

- “Certain English terms can have various French translations. For example the word ‘lid’ can mean 1-lid/cover or 2- eyelid. If your system uses a dictionary which shows the different possible translations and their context, there would perhaps be a way of prioritising to different possible translations according to the context and choosing the best. For example, for the project in question, for segment 10, every translation linked to the terms (eye, ophthalmology, medical) should be given priority over the general translations.”

User 3:

- “I am registered with the portal, but am not able to download or see the tasks: when I click on the name of the document or on ‘Details’, no page is shown.”

User 4:

- “I started to work. Was able to do a couple of sentences. Then it stopped. Every time I tried to start the job again I had an error message on the computer. Then I realized I do not have the Mac version which is required. Unfortunately I am not able to ‘end task’ as required.”

After each post-editing tasks, the user should also spontaneously answer: *Please tell us your sentiment about the task you just finished?*

- “Grrrr, I had to redo it twice to achieve it.”
- “Too hard, output was more or less of no use.”
- “Almost there, I enjoyed it in some ways.”
- “Fully satisfied, output made my life easier.”

The global comments from post-editors were (classified by frequency):

- “Almost there, I enjoyed it in some ways” - *this comment appeared 8 times.*
- “Fully satisfied, output made my life easier” - *this comment appeared twice.*
- “Too hard, output was more or less of no use” – *this comment appeared twice*

The rest didn’t answer this question.

Conclusion

More than 50% of the feedback concerned technical issues: first, users didn't understand all of the functionalities (the user interface seems easy but the process is too complicated). Then, the platform was not always accessible due to technical issues or because of incompatibility with browser or hardware platform.

User sentiment just after post-editing was very positive for this first experiment: more than 80% were satisfied or fully satisfied.

The experiment demonstrates that the bilingual post-editing is user-friendly enough for a community of this kind. It also demonstrates the high potential of success for the ACCEPT project among a community of translators/post-editors.

Post-task Questionnaire for TWB Community

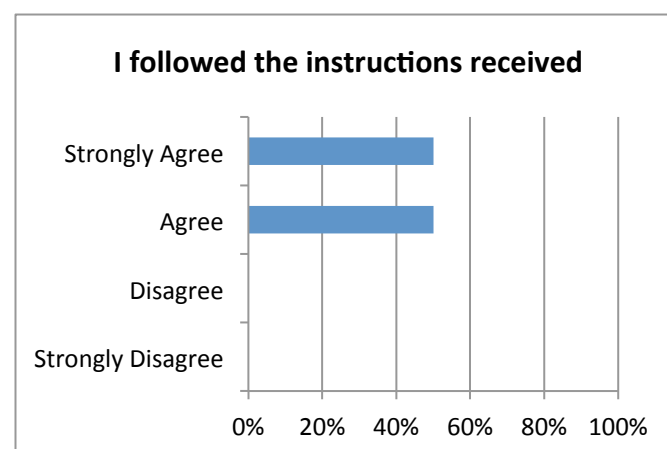
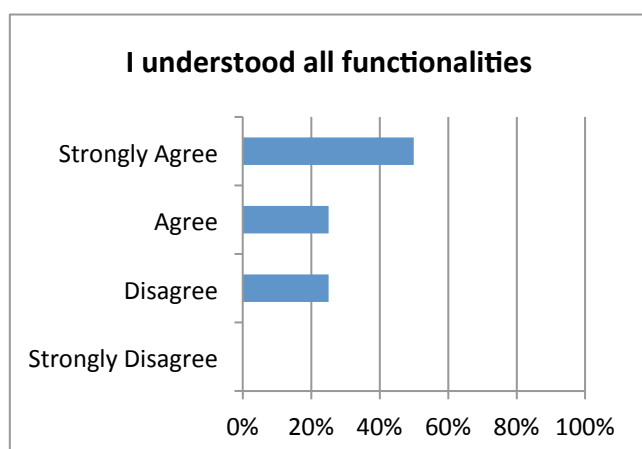
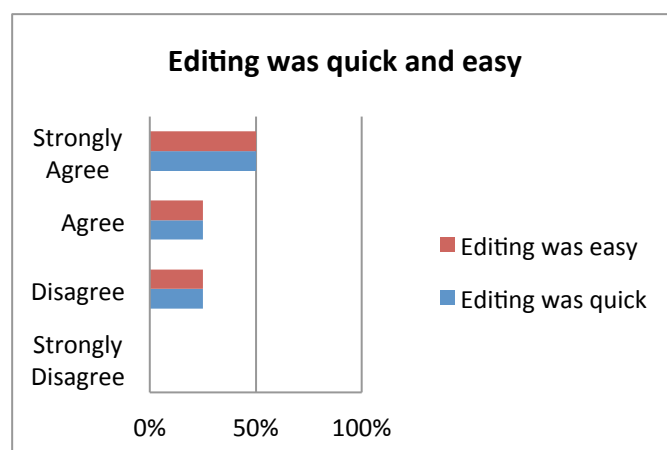
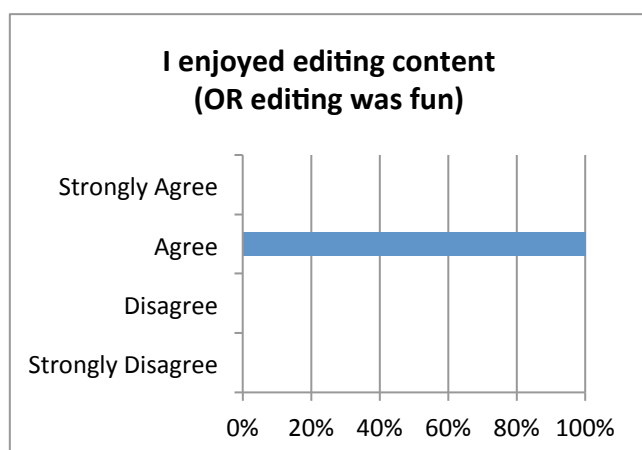
In addition to the immediate query on their assessment of the post-editing task, a new survey was conducted one week later, after the subjects had had time to reflect.

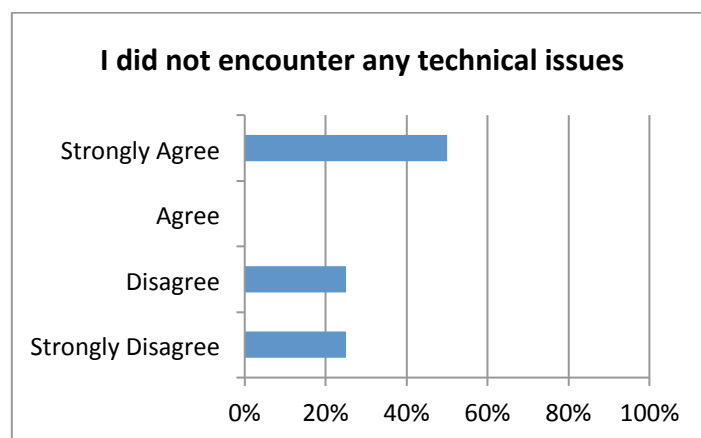
Results of the Post-task Questionnaire

General Evaluation

In this part of the survey, the post-editor was asked to evaluate his/her editing experience and the interface.

User evaluation about his/her editing experience and the interface:





Comments or improvements concerning editing experience:

- “The platform is not using the FR punctuation (like a space before a colon or a semi-colon, « » and not " ", etc.) so there are many corrections to be made.”
- “I would like to be able to download the source document for reference. That would be particularly useful for documents containing tables and diagrams, or just to have a sense of what is part of the same paragraph if there is a doubt...”

Comments or improvements concerning the interface:

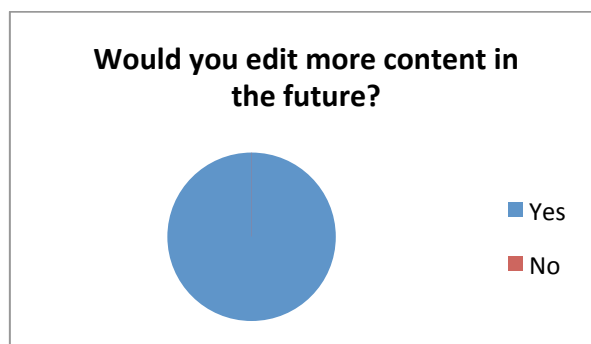
- “Would it be possible to have a functionality to widen the windows where we type our text in? It would be very helpful for larger segments.”
- “It would be useful to be able to find the occurrences of a particular term, in the TM or in the text already translated, to keep consistent.”
- “Resizable windows for source sentences, in edition sentences and Comments.”

Comments or improvements concerning the rules (functionalities) / instructions:

- “I would like to know which tasks are already being edited or have been edited, and which tasks I worked on. I understand that at the moment you need as many edits as possible, even if multiple persons edit the same task but it would be interesting to know if task 1 has been edited 10 times and task 3 none.”

Motivation to Edit

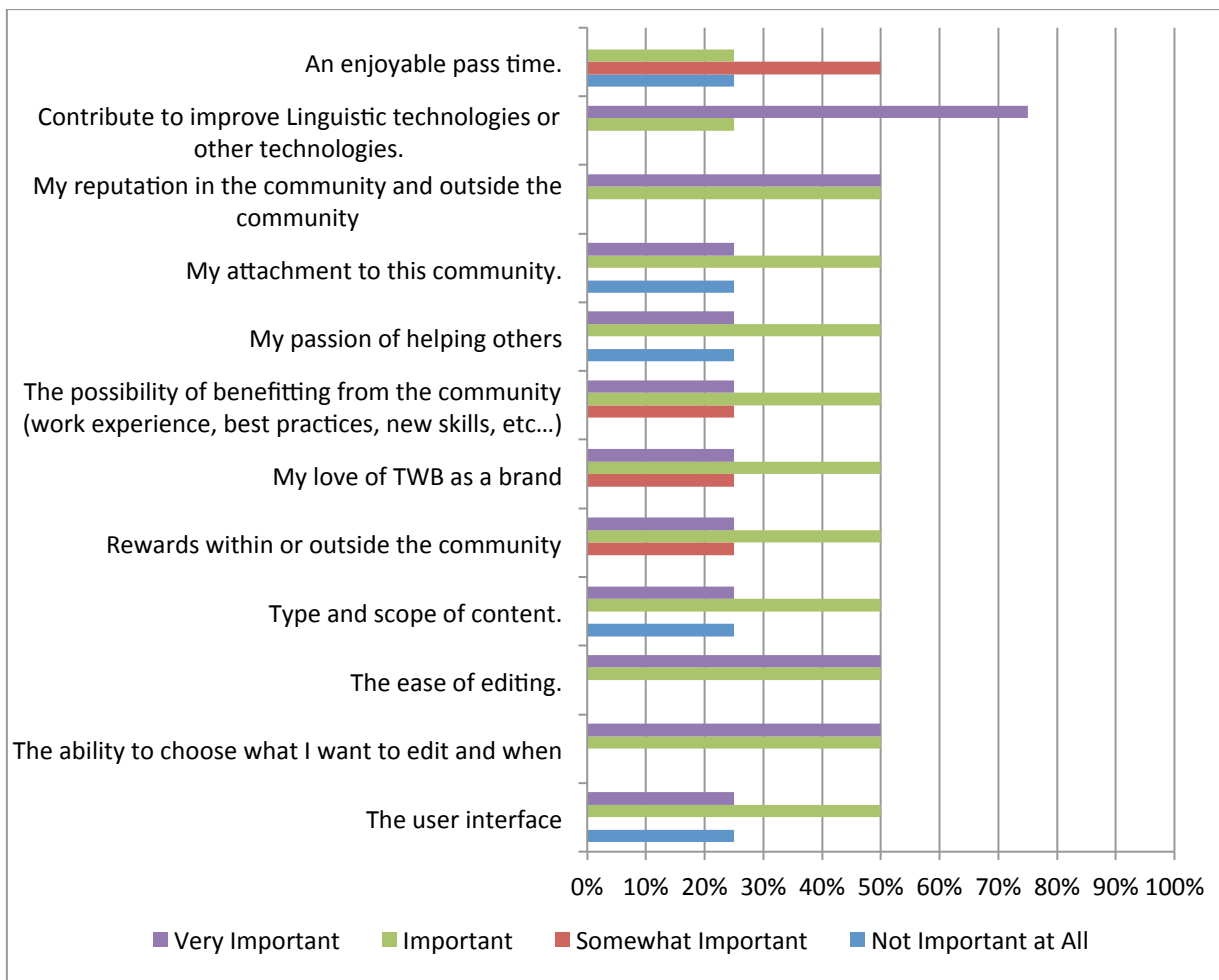
In this part of the survey, the post-editor was asked to evaluate his/her motivations for wanting to continue this experience.



If “Yes”, what is your strongest motivation?

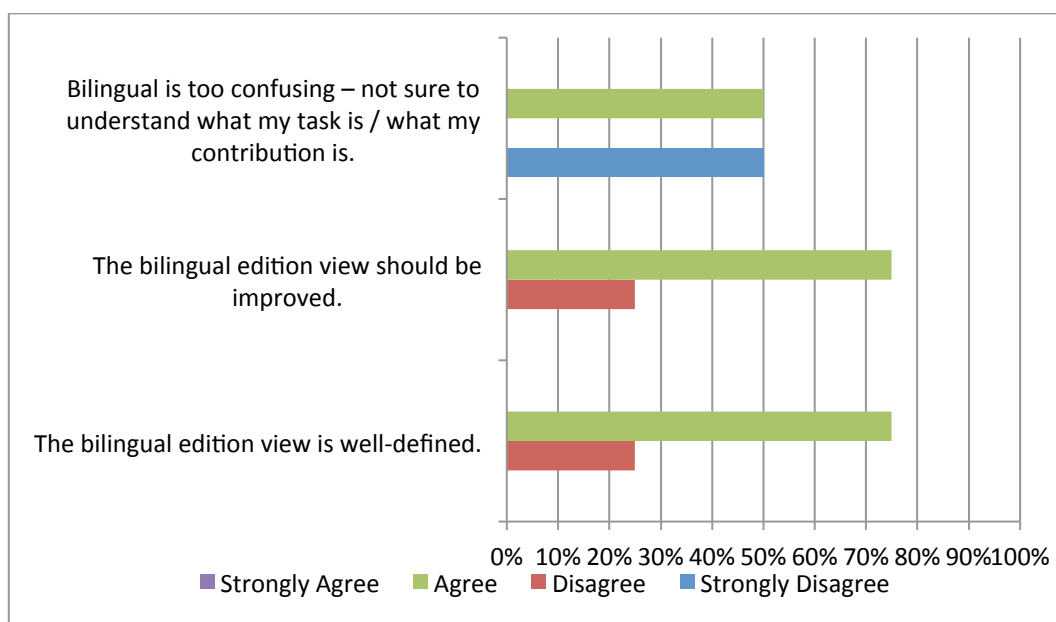
- “The possibility in the future to get a sort of certificate for our participation?”
- “To participate in a research project which can make our work more efficient.”
- “To participate in a European project.”
- “The project seems very promising and I'm interested in helping.”

How important were (or would be in the future) the following aspects in your decision to edit content?

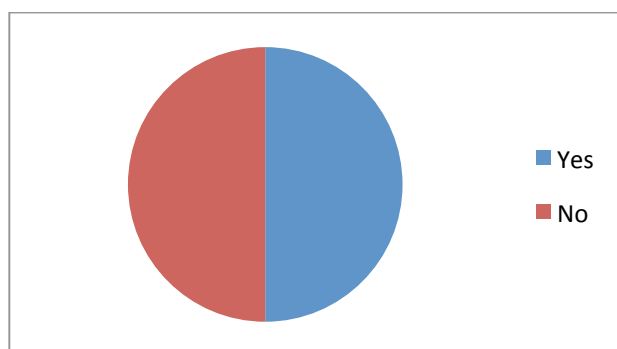


Work Environment Experience

Bilingual edition view :



Do you think it would improve your current work (process and quality) as a TWB volunteer to continue editing content as with ACCEPT experience?



Why?

- “Because I can gather a lot of experience and expertise.”
- “I do not really know I already used translation platforms (3), and it really speeds up the translation process. I think the quality of the machine translation still needs to be improved. It could be a fantastic tool to edit a translation that would be mostly accurate. But right now, it's almost like having to translate from scratch, like in a regular TWB translation, except that we are missing the full context in the form of the complete source document.”

Summary

The questionnaire has given a reasonable amount of initial information about user expectations, the community, the user interface, and the post-editing process in general.

All members of this first experiment enjoyed editing because they found editing was quick and easy, and the instructions given by the community manager were clear. But the user interface is still considered as too complicated, because not all features are sufficiently clear to the users (e.g. validation of a task: why is it necessary and how does it work?). 50% of this community does not think that the TWB community would gain from the deployment of a system of this kind.

All volunteers for this first experiment would like to continue editing in the future for the ACCEPT project because (a) they like the idea of participating in a research/European project, (b) they can edit what and when they want, and finally (c) they appreciate contributing to research in language technology.

The most important aspects motivating them to decide to edit content were their attachment to and reputation in the TWB community, and then the user interface. Rewards within and outside of the community were mentioned by different users.

Evaluation

Symantec Community

Manual Evaluation

After a test-run of the post-editing tasks, it was decided that seven posts per group (e.g. bilingual) were too many because it would have taken too long to post-edit them. Since the participants were volunteers, Symantec wanted to maximise participation and minimise frustration by keeping post-editing time to a minimum. The number of posts per set-up was consequently reduced to 4.

Table 1 displays the number of segments for each set-up and the number of words. The average number of segments for each task was 8 and the average word count was 140 words.

	Segments (DE/FR)	Words (DE/FR)
Bilingual – Task 1	3/4	75
Bilingual – Task 2	5	103
Bilingual – Task 3	12	206
Bilingual – Task 4	8	120

Table 1: Number of segments and words per task

Recruitment was an unexpected challenge; users were not interested in learning about the study or taking part in it, as shown by the lack of replies to the recruitment messages. “Passive” recruiting, i.e. posting about the study publicly for everyone to see, did not result in any users wanting to volunteer in the experiment, despite the fact that the board containing the recruitment post at top position had 82 views on the first day and about 60 views every following day from 09/10 to 23/10 (German forum). The researchers found, however, that once participants did take part, they were enthusiastic about it. This was evident from the posts the participants posted publicly and also from private messages to the project leader. Thus, for future studies, alternative recruitment strategies may be needed to overcome this challenge. It is evident that the number of participants for this pilot study is too low to allow for any interpretation that goes beyond initial indicators.

With regards to the selected texts, the researchers were aiming at selecting similar texts that could be compared across the three set-ups (monolingual, monolingual with options and bilingual). Unfortunately, direct comparability cannot be guaranteed. Thus, an experiment with participants editing the same texts in different set-ups would allow for a more accurate comparison - but would also require more participants.

Technical issues encountered were mainly based around an insufficient explanation of not self-explanatory user interface. Login issues were encountered when users created a user account with a different email address to the one initially indicated. Some of the results could not be recorded accurately, as the users edited the whole task in one segment instead of each segment separately or copied everything into a text editor and back into the online editor because it was more convenient for them. This led to a simplification of the interface and the development of a video which demonstrates how the interface works (these materials will be delivered with WP6 deliverables).

Automatic Metrics

The raw MT output and the post-edited output was also rated using the Meteor 1.4 (Denkowski and Lavie, 2011) and TER (Snover et al. 2006) automatic metrics, comparing them to two sets of reference translations, with one using formal language and one a more informal style, as we wanted to uncover whether the post-edited data is closer to formal or informal language.

MT				Users		
French						
Task	Meteor	TER-1	TER-2	Meteor	TER-1	TER-2
Bilingual	49.5	77.8	66.8	53.7	76.4	65.7
German						
Task	Meteor	TER-1	TER-2	Meteor	TER-1	TER-2
Bilingual	48.3	76.5	66.8	53.0	72.2	64.9

Table 2: Meteor and TER scores

The table above shows Meteor and TER scores obtained by comparing the MT output with both sets of reference translations and by comparing the post-edited data with both sets of reference translations. This was performed for both language pairs. For French, the Meteor scores increase by 4.2 points absolute and the TER scores improve by 1.4 points and 1.1 points absolute. For German, the scores show that bilingual post-editing leads to improved scores compared to the raw MT output. The Meteor scores increase by 4.7 points absolute and the TER scores improve by 4.3 points and 1.9 points absolute.

The improvements are minor, with degradations of scores also being evident from MT output. However, some similarities between the human scores and the automatic metrics can be identified. For German, fidelity has the highest percentage of improved segments for the bilingual set-up, which is reflected by all automatic metrics.

The following figures display the post-editor productivity for the bilingual set-up and the two languages. This is broken down into editing time and average number of keys pressed per task. As

can be seen when comparing Figure 2, the German participants spent a considerably longer period of time on a task. One French user spent more time on average per tasks, which is close to the data of the German participants. This general trend of French participants spending less time than German participants is also mirrored by their number of keystrokes pressed. Combined for all German users, there were a larger number of keys pressed per task on average than for the French users. The difference between the number of key presses is, however, smaller than the difference between the average editing time spent per task.

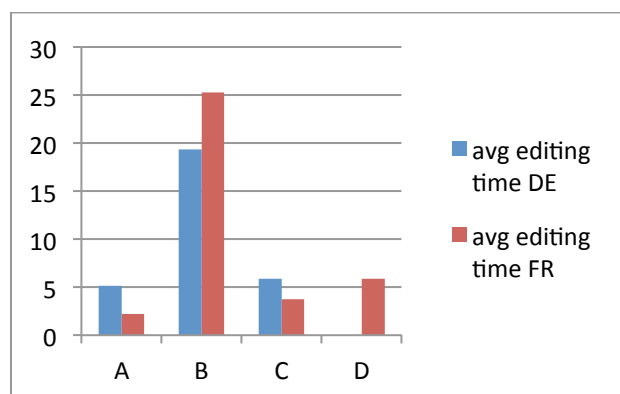


Figure 1: Average editing time for German users in minutes

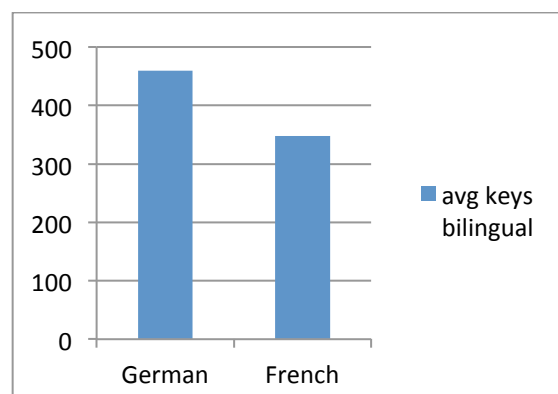


Figure 2: Average number of keys pressed per language

TWB Community

During this first experiment, about 6 500 words were post-edited:

In order to measure the quality of the post-editing tasks, a manual evaluation was performed and automatic metrics were also used.

Manual Evaluation

For the manual evaluation, a comparison between the Machine Translation output and the reference (human translation/post-editing content) was performed. Errors were classified into three categories: accuracy, terminology and language.

The table below summarizes the types of errors found:

Terminology	14%
Accuracy	25%
Language	62%

Table 3: Types of errors found

The main errors in the MT output concern:

- (a) **Noun phrase issues:** e.g. Reservoir control

MT output: Réservoir de contrôle

Post-editing output: Réservoir de contrôle (it should be: « Contrôle de réservoir »)

- (b) **Gender and number agreement issues:** e.g. Diseases that are transmitted this way include airborne diseases and sexually transmitted infections.

MT output: Les maladies qui sont transmis cette manière inclure les maladies et les infections sexuellement transmissibles

Post-editing output: Les maladies transmises de cette manière comprennent les maladies transmises par l'air et les infections sexuellement transmissibles.

- (c) **Noun-adjective inversion :** e.g. active artificial immunisation

MT output: active d'immunisation artificielle

Post-editing output: immunisation active artificielle

Table 4: Detailed report of the error types for each task

ERROR TYPE	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	All tasks
T - Inconsistency	1	1	1	0	1	0	1	0	0	0	5
T - Inappropriate to context	0	0	0	2	0	1	1	0	0	0	4
T - Sector-specific terminology	2	1	0	2	0	0	0	0	0	0	5
A - Omission/Addition	1	1	1	0	0	1	1	1	0	1	7
A - Untranslated text	1	1	0	0	2	1	2	1	2	4	14
A - Incorrect meaning	0	0	0	1	0	1	1	1	0	0	4
L - Grammar/Syntax	3	7	11	10	7	6	3	7	3	4	61
L - Punctuation	0	0	0	0	0	0	0	1	0	0	1
L - Spelling/Typo	0	0	0	0	0	0	0	1	0	0	1
TOTAL	8	11	13	15	10	10	9	12	5	9	102

Automatic Metrics

To measure automatically the quality of the Machine Translation output and post-editing, we used the TER metric (Snover et al. 2006). We compared (a) REF/PE: the reference (the human translation) with the post-editing output, (b) REF/MT: the reference with the Machine Translation output, (c) MT/PE: the Machine Translation output and the post-editing output.

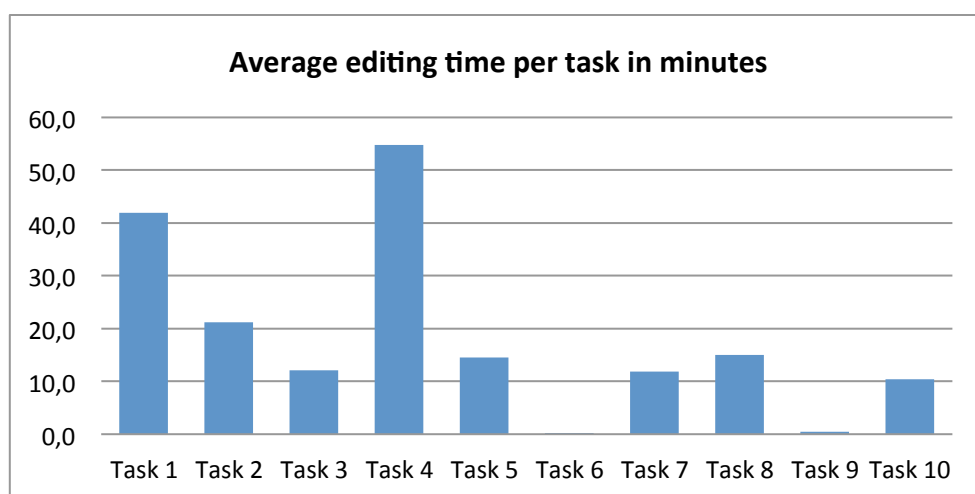
	REF/PE	REF/MT	MT/PE
Task 1	0,24	0,73	0,65
Task 2	0,46	0,71	0,93
Task 3	0,59	0,62	0,59
Task 4	0,27	0,61	0,67
Task 5	0,61	0,57	0,89
Task 6	0,42	0,72	0,57
Task 7	0,62	0,90	0,54
Task 8	0,01	0,50	0,51
Task 9	0,02	0,60	0,62
Task 10	0,56	0,67	0,77

Table 5: Results about the quality of the Machine Translation output and post-editing

For Tasks 8 and 9, we noticed that the reference and post-editing output are very close (if not identical): an investigation will be carried out in order to understand how it is possible to have such extreme similarity between texts. It is possible that the post-editor in charge of these two tasks copied/pasted reference text that he found on the Internet. These tasks are excluded from the study when calculating the following average score.

For this first experiment, the reference and the post-editing output are very close. There is an average TER of 0.47.

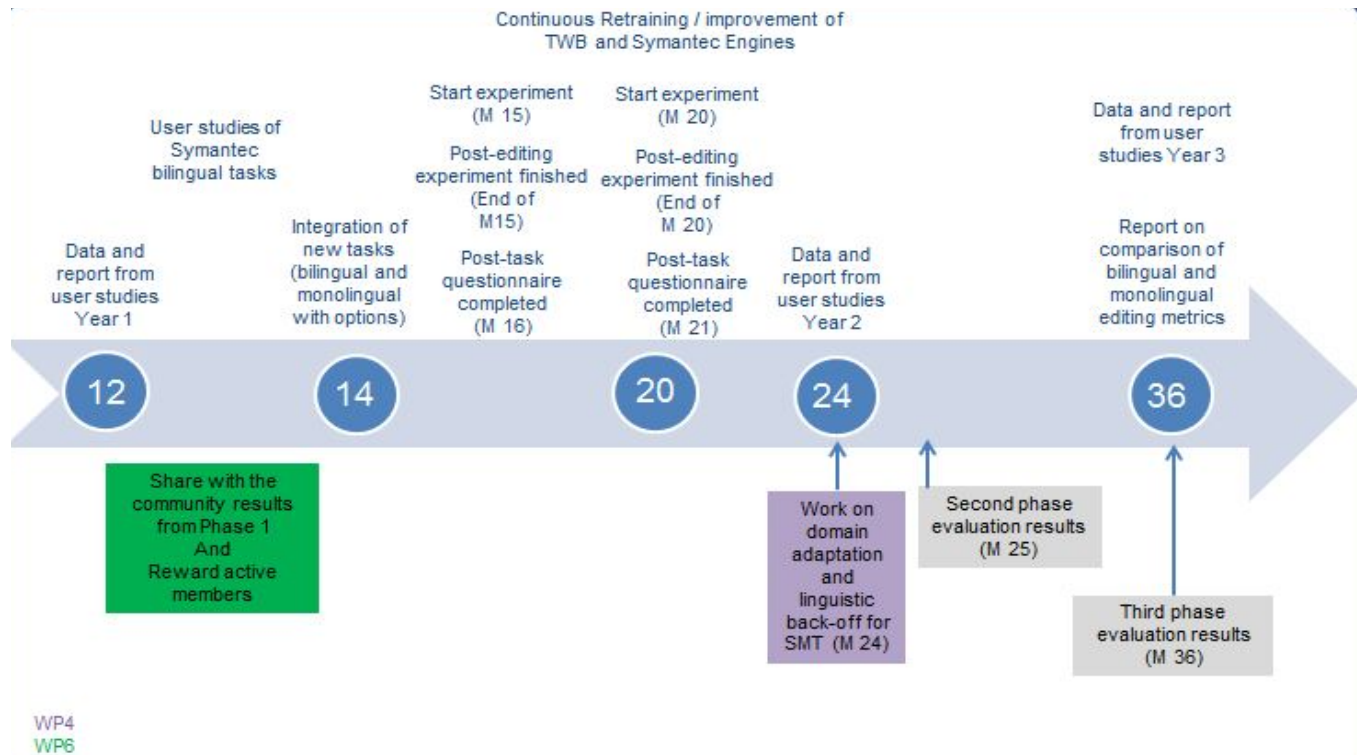
Another important metric is the time spent post-editing during each task:



For this first experiment, even if we had most of the time a REF/MT TER >50% (which means that the human translation and the Machine Translation output are quite distant), we obtained an average of about 22 minutes per task of 500 words. We can also notice that there is quite a large range of productivity for the same volume of words. There are two possibilities: the further the post-editors

progressed in their assigned tasks, the quicker they were, or the difficulty from one task to the other may be varying.

Timeline



The next TWB experiment will be launched in Month 15. The same methodology will be applied. However, we will add the “options” functionality (the post-editors will have multiple choices when post-editing). We will also experiment on some pre-edited content, still for the combination English->French. The post-task questionnaire will change and will be updated according to the specificities of this second experiment.

During Month 15 we will also launch the first experiment for the combination French>English, without translation options and without pre-editing rules.

At the same time, we will share all user data with the owner of WP4 to improve the machine translation for TWB baseline.

Another experiment will be launched at Month 20 but then it will be French>English post-editing with pre-editing rules and with translation options. We will also test the pre-editing rules for English>French during the same period of time.

For Symantec, we will carefully examine which information to display (based on the data collected in the experiment described here) (Month 12-16) before defining requirements for the next user study. We plan for an experiment in Month 21-22.