



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° 7

Name: Monolingual Postediting

Deliverable n° 7.1.1

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: 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

Objectives of the Deliverable	3
Monolingual Post-Editing Experiments at Symantec	3
Summary	3
Recruitment	3
Set-Up	3
Texts Selected	4
Process	5
Results	6
Post-editor productivity	7
Challenges	8
Monolingual Post-editing Experiments at TWB	9
Methodology for the monolingual post-editing.....	9
Pre-task questionnaire results for TWB community	9
Results of the pre-task questionnaire	9
Conclusion	11
Post-task questionnaire for TWB community	12
Results of the post-task questionnaire	12
General Evaluation	12
Motivation to Edit	13
Conclusion	15
Timeline	16

Data and report from user studies - Year 1

Objectives of the Deliverable

The objective of workpackage WP7 is to collect user edits from users who can perform monolingual post-editing in a volunteer or community collaboration context.

The main goal of this deliverable is to provide report about the two first user studies and a first revision of the expected requirements and specifications (timeline, etc.).

The two studies of this work package were done in parallel by Symantec and TWB with their own community and are reported in Sections ‘Methodology for the Monolingual Post-Editing at Symantec’ and ‘Monolingual Post-editing Experiments at TWB’, respectively. The Machine Translation systems used in this study are the baseline MT systems developed as part of WP4 ([see Deliverable D 4.1](#)).

Monolingual Post-Editing Experiments at Symantec

Summary

We carried out a machine-translation post-editing pilot study with users of a technical support forum. For both language pairs (EN-DE, EN-FR), 4 native speakers of the target language were recruited. They performed monolingual post-editing tasks (4 tasks) and assisted monolingual post-editing tasks (4 tasks) on machine-translated forum content, as described in this document, as well as bilingual post-editing (4 tasks) (see WP8). The post-edited content was evaluated using automatic metrics (Meteor, TER). We found that monolingual post-editing can lead to improved scores, although scores improved considerably more for the bilingual set-up (see WP8).

Recruitment

Users were recruited within the forums by the forum administrators. Users targeted were native speakers of the target language. Thus, four users for German and five users for French were recruited initially. For French, 28 ‘gurus’ or active users were invited via private message. Out of these, 17 did not reply, five users declined and six 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. Two users did not reply and two users declined; eight users accepted the invitation to take part in the study, out of which four had to be discarded because they did not fit the profile (e.g. their first language was not German). The final recruitment success rate was 33% for German and 15% for French.

Set-Up

The machine translation system used in this study was trained on bilingual data both from in-domain data, e.g. product manuals, and out-of-domain data, i.e. WMT12 releases of EUROPARL and news commentary (EN-DE, EN-FR) using Moses (Koehn et al., 2007). The texts used for post-editing were taken from the English-speaking support forum. They consist 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. It was believed to be disadvantageous for the participants of the study to edit each post three times (monolingually, monolingually with translation options and bilingually) since they would learn from each post, which is why a method of clustering similar posts together was deployed. To create datasets of similar size and content, features of the texts, such as the type-token ratio and length of the post were considered. The first set-up involved only the raw MT output that was presented to the users, while the second set-up involved additional translation options that were presented with the raw MT output. Translation options are obtained directly from a [Moses server](#) via XML-RPC, an example of which is shown by the following code snippet:

```
import xmlrpclib
server_ip = '' # IP of mosesserver
server_port = '' # Port on which mosesserver is listening
server_proxy = xmlrpclib.ServerProxy('http://server_ip:server_port/RPC2')
params = {'text': 'this is a test', 'topt': True}
trans_results = server_proxy.translate(params)
print [x for x in trans_results['topt']]
```

Figure 1: Retrieving Translation Options from Moses server

Since a large number of translation options can be retrieved, filtering these options seems necessary to avoid overwhelming the post-editors. The following basic filtering approach was used:

For each token in a tokenized translated string:

- a) Find source alignment information (start and end index) for the current token. If no alignment information is available, skip to next token.
- b) Find translation options whose start index corresponds to the start or end index identified in the previous step.
- c) Ignore phrases that contain multiple tokens (e.g. phrases that contain a space).
- d) Ignore phrases whose fscore is below a certain threshold (e.g. -8).
- e) Remove “duplicated” phrases (e.g. those have the same value but different scores).
- f) Keep a certain number of options per token (e.g. 5).

Texts Selected

After a test-run of the post-editing tasks, it was decided that seven posts per group were too many because it would have taken too long to post-edit them. Since the participants were volunteers, we wanted to maximise participations and minimise frustration by keeping the post-editing time to a minimum. Thus, the number of posts per set-up was reduced to four. Table 1 displays the number of segments for each set-up and the number of words.

Task	Segments (DE/FR)	Words (DE/FR)
Monolingual – Task 1	9	114
Monolingual – Task 2	7	101
Monolingual – Task 3	11	235
Monolingual – Task 4	7	108
Options – Task 1	10	137
Options – Task 2	6	57
Options – Task 3	11 / 12	298 / 206
Options – Task 4	8	75 / 120

Table 1: Number of Segments and words for each task

Process

The users performed the post-editing tasks using a portal that was especially developed for post-editing (URL), the interface of which is displayed in Figure 2 (see Deliverable D5.3). The left half of the window shows the full text to be edited for that particular task. In the top right edit box the user can edit the current segment. Comments can be made in the edit box at the bottom right. All edits were saved automatically. During the post-editing process, editing time, keystrokes, usage of translation options etc. were recorded in the portal.

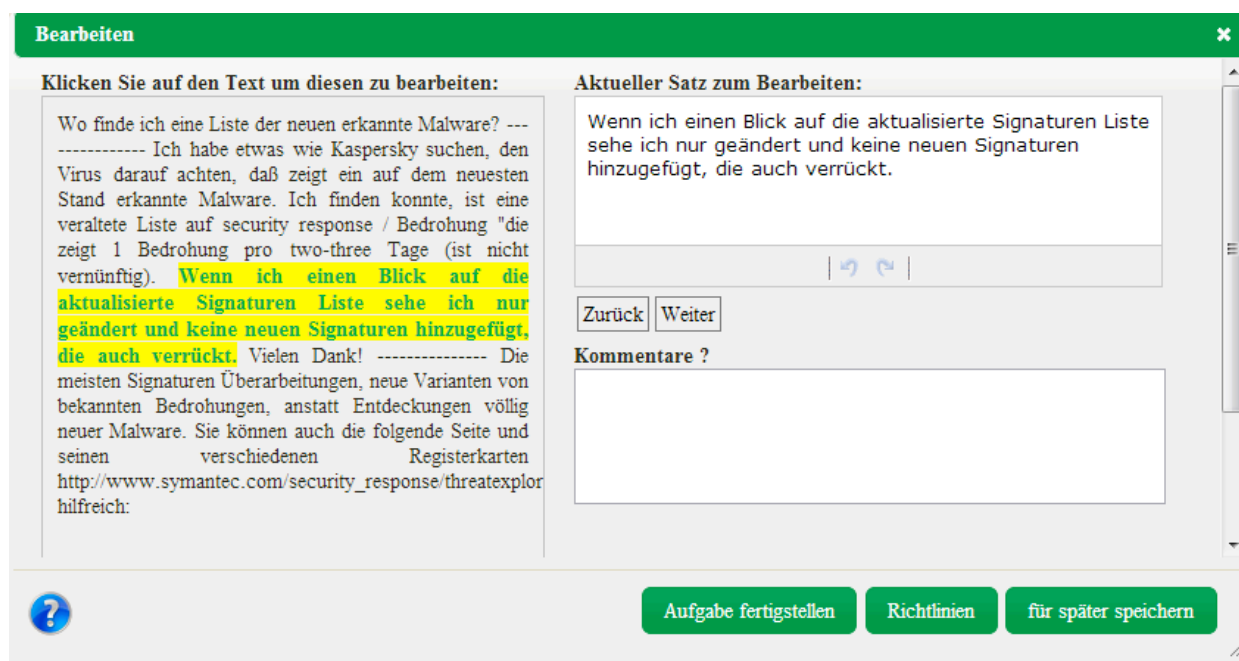


Figure 2: PE Interface - monolingual

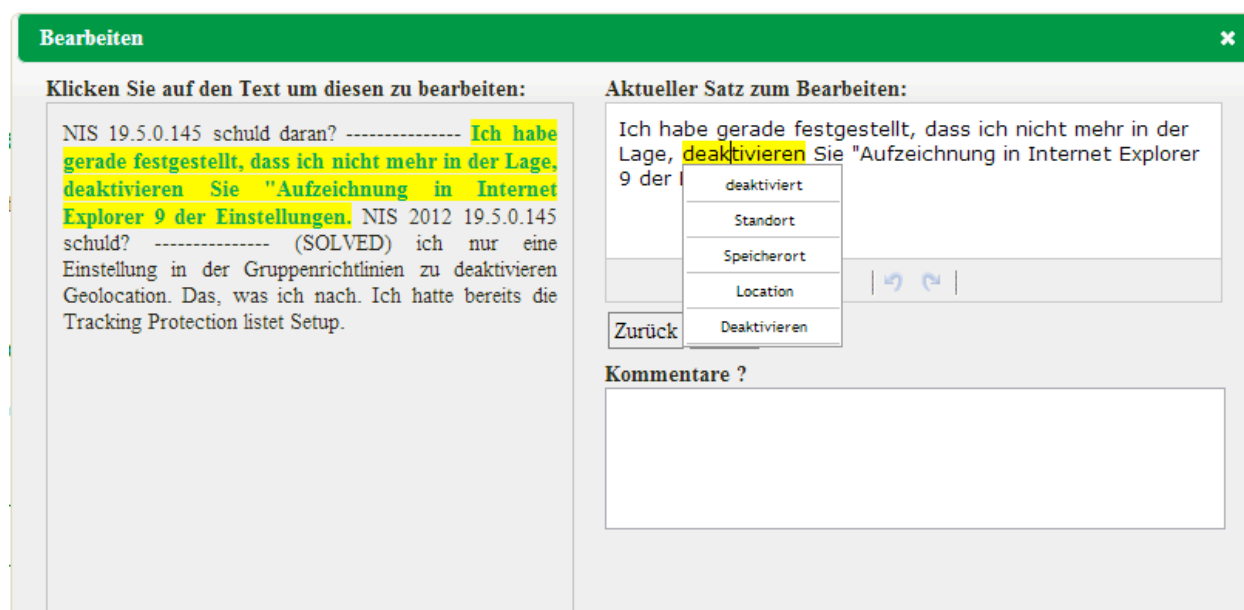


Figure 3: PE interface- monolingual with options

Results

In order to be able to use automatic metrics to score the post-edited content, we created two sets of reference translations for each of the texts. One set was written in a formal style (TER-1) and the other in an informal style (TER-2). 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 determine whether the post-edited data is closer to formal or informal language. Mt vs PE, PE vs. Ref

	MT			USERS		
French						
Task	Meteor	TER-1	TER-2	Meteor	TER-1	TER-2
Monolingual	49.8	79.7	76.4	50.0	77.7	76.8
Options	52.0	74.1	76.8	51.5	71.9	77.3
German						
Task	Meteor	TER-1	TER-2	Meteor	TER-1	TER-2
Monolingual	48.9	70.4	66.9	45.7	74.5	72.5
Options	47.0	71.3	69.5	46.3	74.1	71.2

Table 2: Automatic Metrics Scores

Table 2 shows Meteor and TER scores that were 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. It is evident that the difference between the two set-ups, monolingual and monolingual with options, is minor for both language pairs. For French, it seems to be the case that the set-up with translation options performs slightly better, whereas it is the monolingual without options for German. It has to be considered, however, that the options were not used very much, as can be seen in Figure 7.

When comparing the post-edited scores with the MT scores, it seems that for French, the users performed slightly better in both set-ups when considering the Meter scores and TER-1. For TER-2 (set of informal reference translations), both scores are lower than the ones for the raw MT output. For German, however, it seems that the users perform worse than the raw MT output, which is reflected in all scores. This suggests that automatic metrics might not be able to reflect the post-editing results truly. Thus, we will include comprehensive human evaluation for future studies.

Post-editor productivity

The following figures display post-editor productivity across the two set-ups 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 4 with Figure 5, the German participants spent a considerable longer period of time on a task for both set-ups in general. One user for French 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 the number of keystrokes. Combined for all users for German, there was a considerably bigger number of keys pressed per task on average than for the French users.

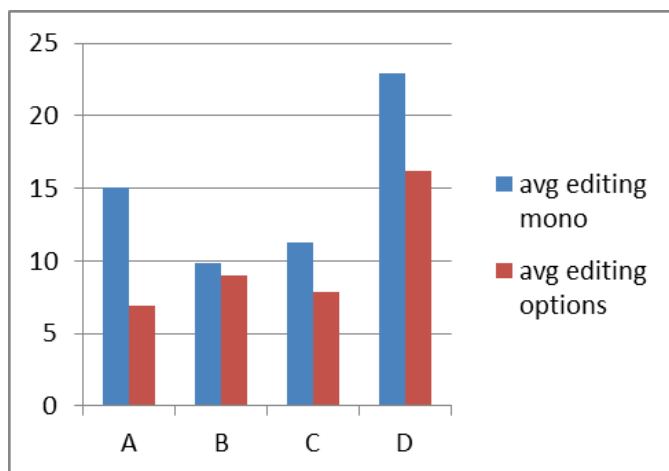


Figure 4. Average editing time for German users in minutes

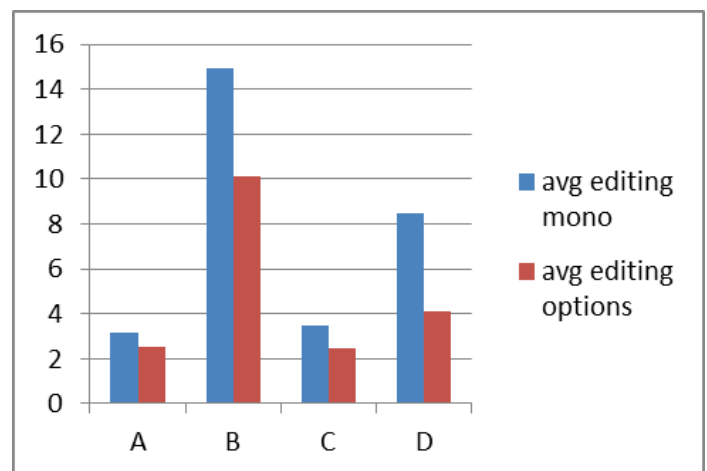


Figure 5. Average editing time for French users in minutes

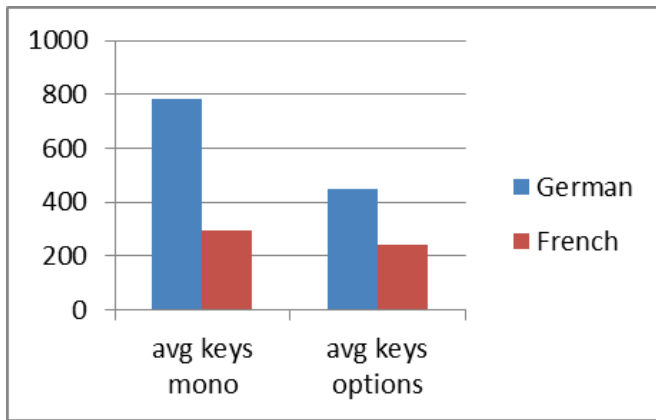


Figure 6. Average number of keys pressed per set-up

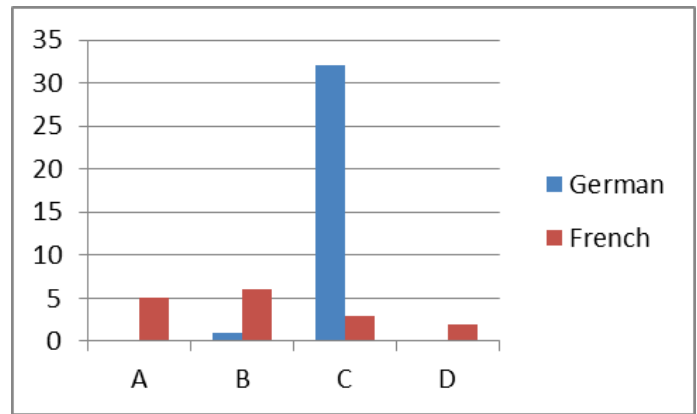


Figure 7. Absolute number of translation options used in total per user

Challenges

Recruitment was an unexpected challenge; users were not interested in learning about the study or taking part in it, as demonstrated 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 which was evident in posts the participants posted publicly and 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. Nonetheless, the experiment suggests that monolingual post-editing is not an unrealistic exercise, assuming forum users, for example, are willing to engage in it.

With regards to the texts selected, 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 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 they 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 a video being developed, which demonstrates how the interface works.

Monolingual Post-editing Experiments at TWB

Methodology for the monolingual post-editing

The TWB documents to be post-edited were translated automatically by 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) and other sources. For more details, please refer to Deliverable 4.1.

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 monolingually thanks to the post-editing environment (developed in WP5). 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.

During this first experiment, about 4 000 words were post-edited for monolingual tasks. No evaluation was performed. In the next sections, we describe the results of the pre-task and post-task questionnaires.

Pre-task questionnaire results for TWB community

The pre-task questionnaire was carried out by the whole TWB community for the ACCEPT project and consequently the results below concerned both monolingual and bilingual post-editors. Since this questionnaire was anonymous, we were not able to distinguish between the two groups of post-editors and the results are therefore included both in this deliverable and Deliverable D8.1.1

Results of the pre-task questionnaire





How old are you?

under 18		0.0%
18-24		0.0%
25-30	<div></div>	16.7%
31-40	<div></div>	22.2%
41-50	<div></div>	33.3%
over 50	<div></div>	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%

Feedback during post-editing for TWB community: User Inputs

During the post-editing tasks, spontaneous feedbacks from post-editors were collected by the community manager:

User 1:

"Ce qui est sûr c'est que pour moi le monolingue est quasi mission impossible (surtout sur des textes médicaux j'aurais trop peur de mal interpréter, on a vraiment besoin du document source pour s'en sortir). Pour moi j'avoue que le projet en monolingue est particulièrement difficile. Je ne relis presque jamais des textes sans voir la version source, je ne me sentirai pas capable d'interpréter dans ce type de cas."

(English translation: "What's for sure is that for me the monolingual is almost mission impossible (especially for medical texts i was scared about misinterpreting, we really need the source document in order to do a good job). I admit that for me the monolingual is particularly difficult. I hardly ever edit texts without having seen the source version, I don't feel capable to interpret in these circumstances.")

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

Volunteers for this 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.

Monolingual edition view is considered as very complicated for users, and they requested to have the source text to edit. As the content is considered as sensitive (medical field, content to be translated for ONG), they need to be confident with the text they edit if there is no source text.

Post-task questionnaire for TWB community

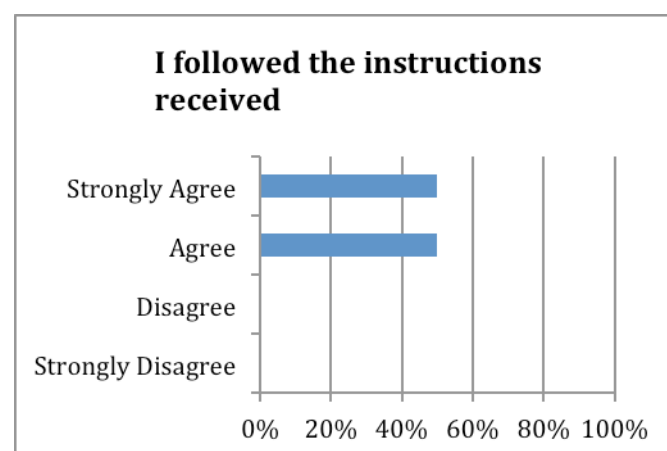
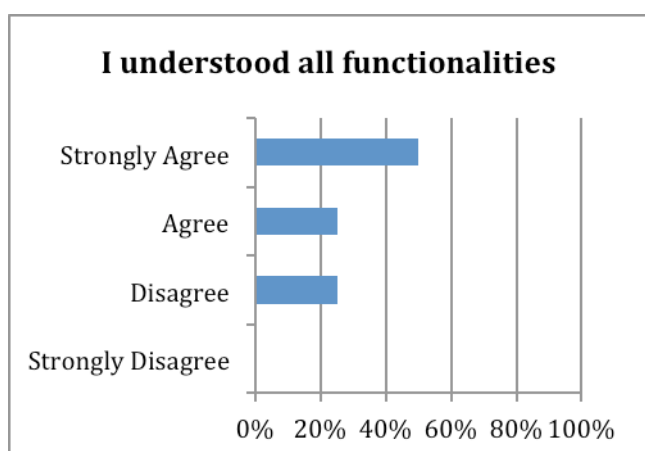
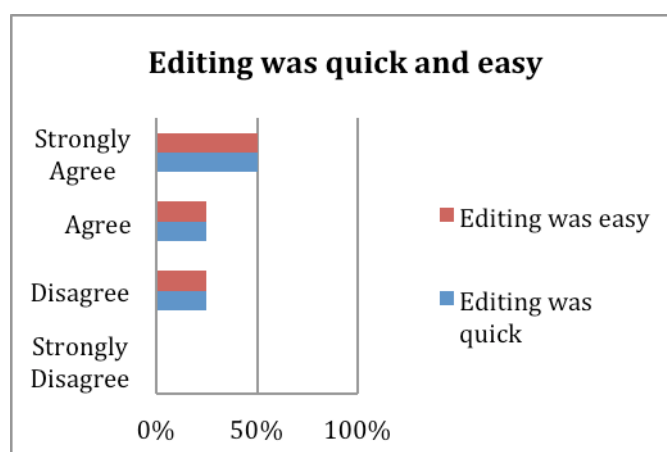
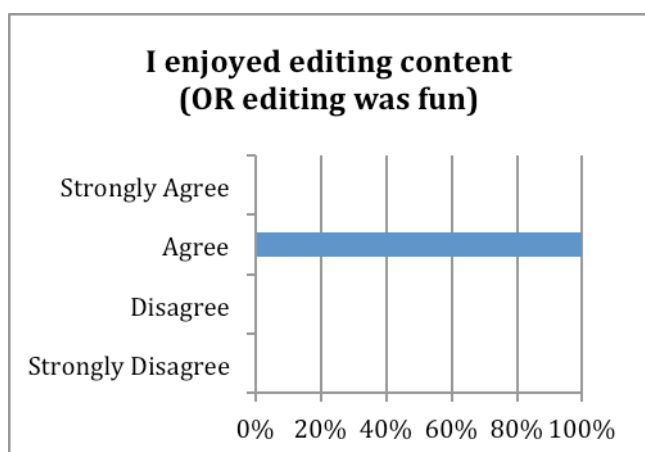
In addition to the immediate query on their assessment of the post-editing task, a new survey was conducted one week later, after having had time to reflect.

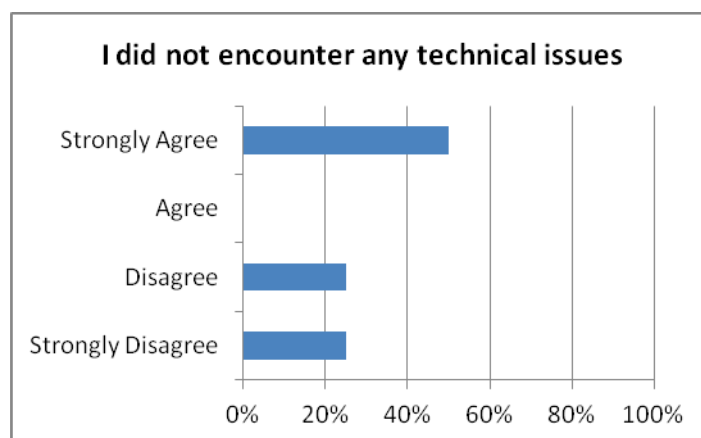
Results of the post-task questionnaire

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 the editing experience:

- “Please provide the original text every time. My second project was a monolingual post editing. I managed to figure out the original texts, but it would save up time if they could also be provided. Thanks!”

Comments or improvements concerning the interface:

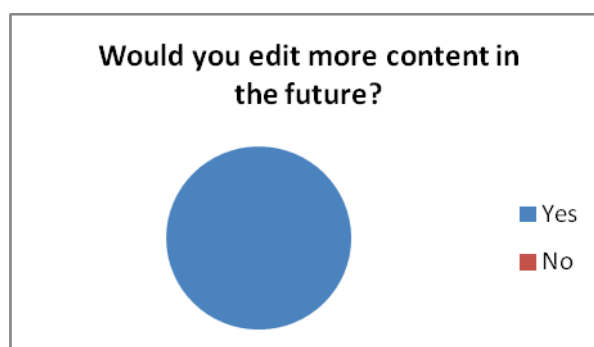
- “Would it be possible to have a functionality to widen the windows where we type our text in? It would be very helpful for larger segments.”
- “It would be useful to be able to find the occurrences of a particular term, in the TM or in the text already translated, to keep things consistent.”
- “Resizable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

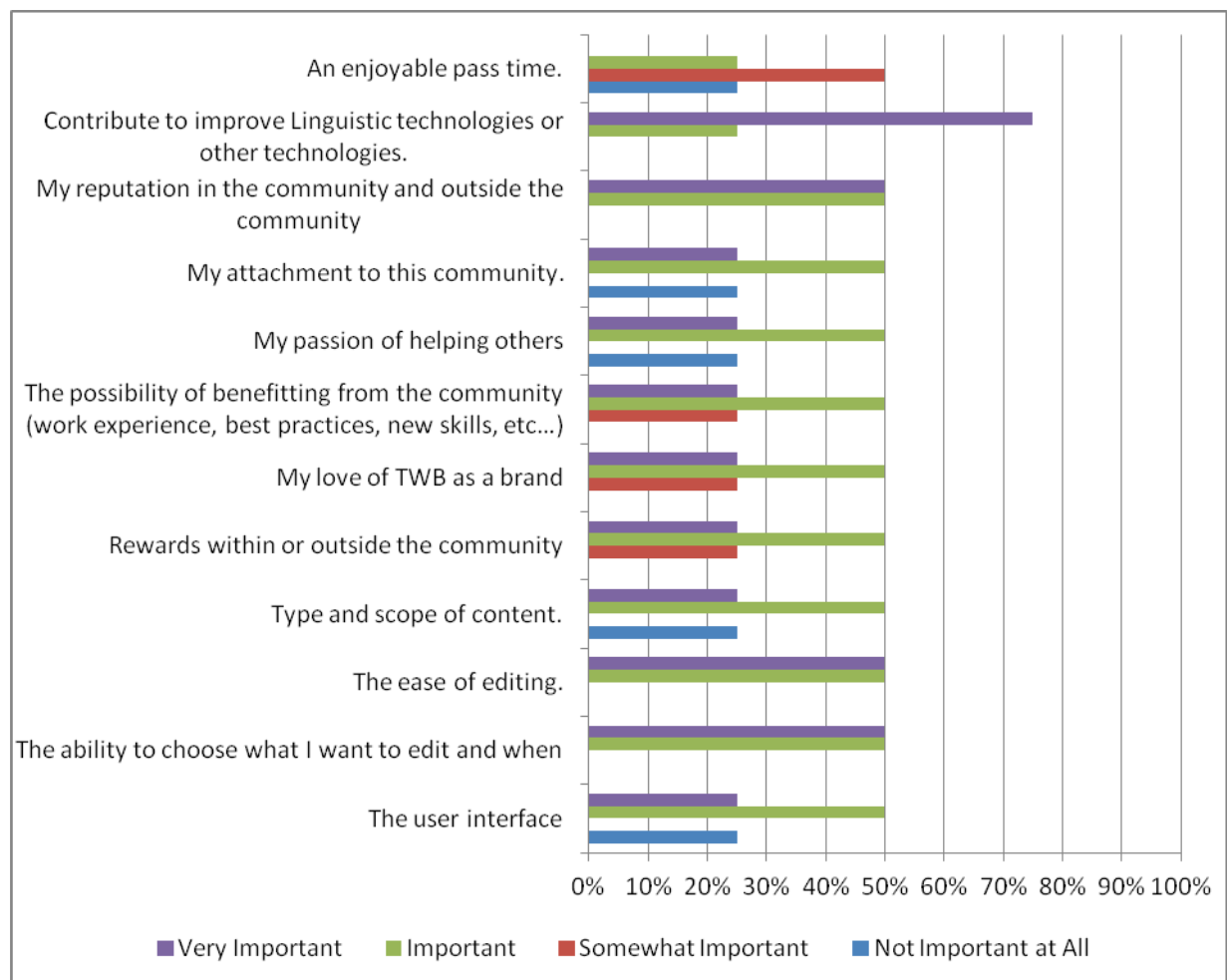
In this part of the survey, the post-editor was asked to evaluate his/her motivations for wanting to continue this experience.



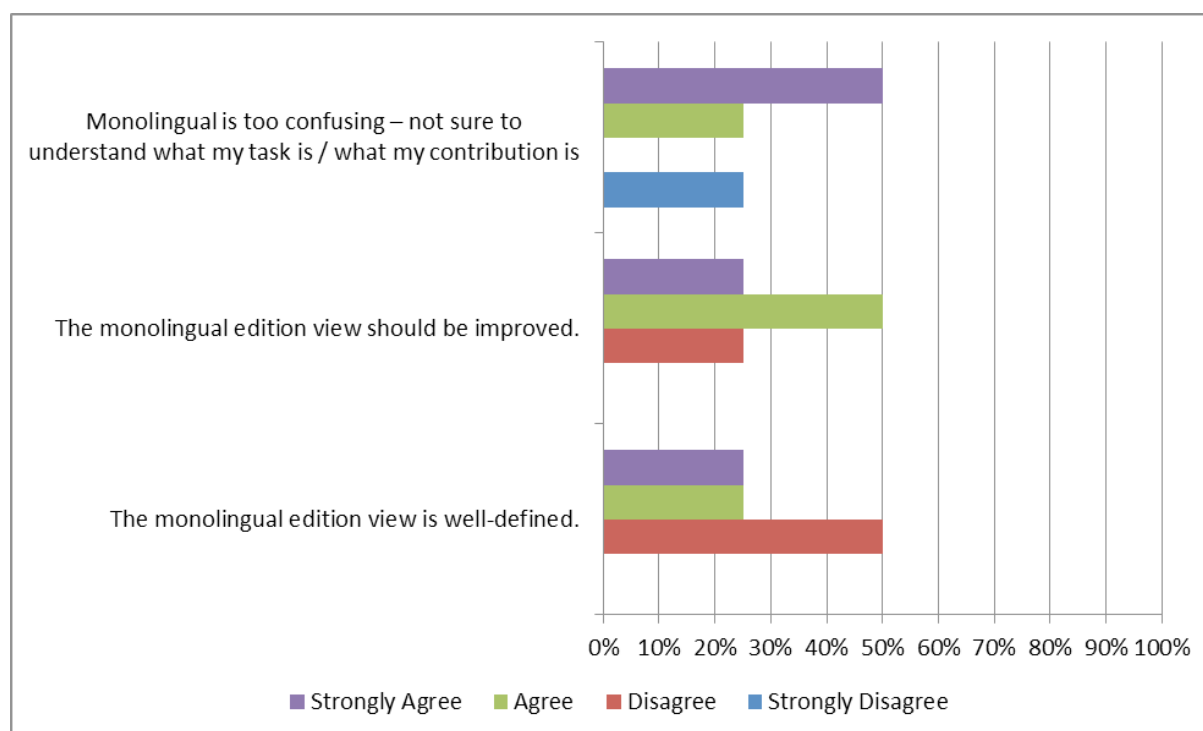
If “Yes”, what is your strongest motivation?

- “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 / Monolingual edition view:



We notice also that the monolingual edition view should be improved: more than 50% of the users didn't like this view and approved the bilingual view.

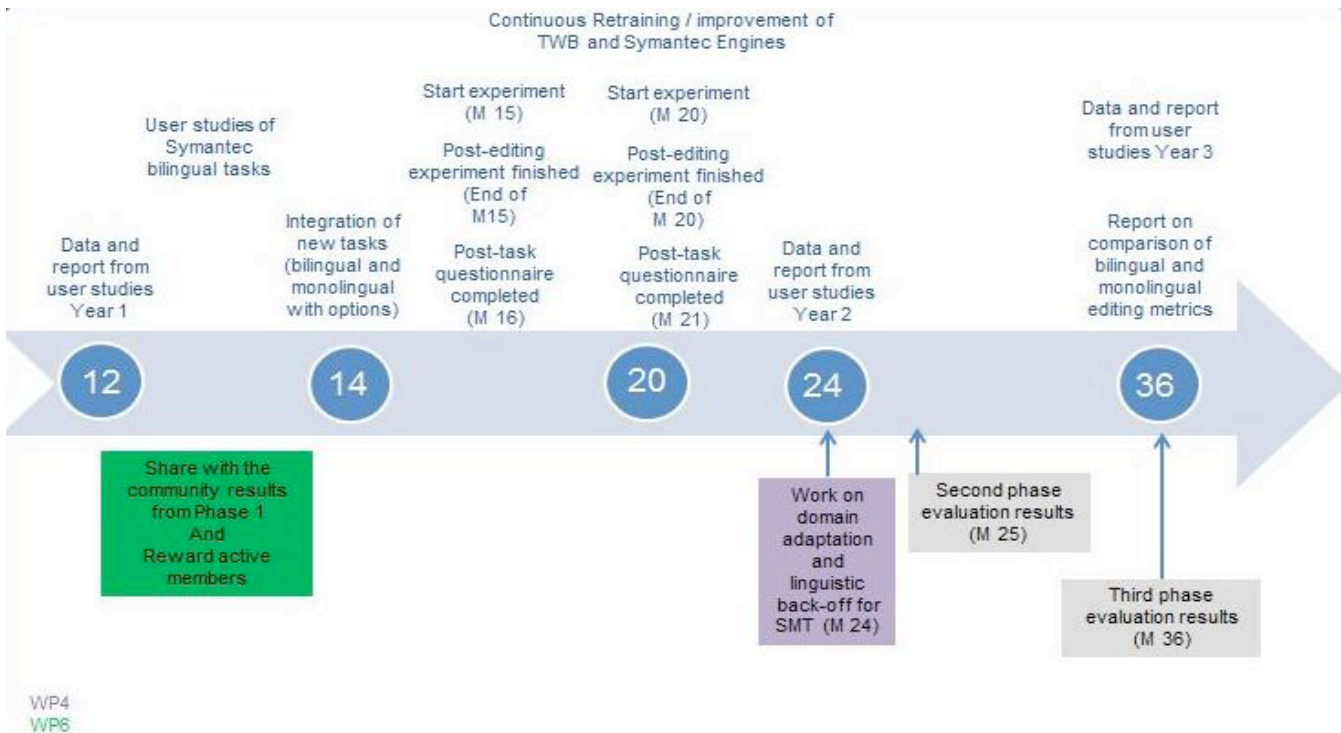
Conclusion

This questionnaire gave us quite a lot of information about this experiment, in terms of user expectations, the community, the user interface, and the post-editing itself.

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, partly because some features are not very clear for the users (e.g. validation of a task: why is it necessary and how does it work?) So 50% of this community does not think that it will help them to give more to the TWB community to have such a system deployed and available.

The monolingual view should be improved for this community. Only 4 000 words were post-edited because more than 50% of the users didn't like this edition and preferred the bilingual view.

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.