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

Seminar Material on Post-Editing – Edition 2

Workpackage n° 6	Name: Community Development
Deliverable n° 6.2.2	Name: Seminar Material on Post-Editing – Edition 2
Due date: 31 December 2012	Submission date: 21 December 2012

Dissemination level: PU

Organisation name of lead contractor for this deliverable: Symantec Limited

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

Seminar Material on Post-Editing – Edition 2

Introduction

Our intention is to assist the post-editor who is faced with a machine translated post or sentence, so that they can easily improve the comprehensibility of that sentence or post. Many of these post-editors will have domain knowledge, but may have little experience in translation and may not be conversant with the source language. The language used by communities is characterised by its colloquial nature and technical focus. Grammar, spelling and style are sacrificed for speed. The goal of the post-editing process is to further increase the comprehensibility of the content in the target language, and possibly to adjust its style to that of the target audience. We can expect a range of domain ability, a range of bi-lingual competence and a range of writing ability within the population.

The project intends to provide technology and services to a wide range of users. These users will fall into three general classes:

- Third parties as yet unknown to the project;
- European citizens and their information providers;
- The ACCEPT partners themselves.

Third parties would include a range of individuals including but not limited to self-help groups with translation needs, through non-government organisations and charities, to commercial enterprises, who often have large amounts of information available in the source language(s) they use, but cannot readily provide translated materials in other languages.

Ordinary European citizens will act as the end recipients of these services, in that an increased amount of information will be available in a greater range of languages. Information providers in general, including the commission itself, will have an opportunity to leverage the technology as it develops.

Target Groups, Potential Partners and Other Stakeholders

Symantec is an active member of the Centre for Next Generation Localisation (CNGL) based in Dublin (Ireland). Our association with other industry partners in this organization gives us a platform to attract forum members to participate in the ACCEPT project.

Through the Special Interest Group (SIG) we will test the portal post-editing functionality and its component software elements. The membership of the SIG will consist of technically savvy institutions, both commercial and non-profit, who wish to test the technology and deploy it on their own social software stack, and smaller groups or companies who opt to use the portal to test the technology. The feedback from these diverse groups will serve to guide the latter stages of development of the project. We expect to grow the influence of this SIG from a few members in the first year to a group of informed and supportive members in the final year.

Challenge of Post-Editing

Post-editing machine translated content is rapidly becoming an industrial process, in that, with the growth of MT usage in the translation industry, the need for such services is growing. However, the process is not well understood, the tools are limited, and, outside the use cases where the post-editing process is conducted by professional translators, relatively little work has been published.

Post-Editing Assistance

Description

The project will develop post-editing assistance functionality for machine translated content in German and French. This functionality will be based on the work started in CAITRA and continued in CASMACAT based in the University of Edinburgh. The initial study in the area will establish a baseline, based on current industrial practice, on the ACCEPT portal with limited functional assistance (e.g. using multiple translation options). There are two aspects to the work: first, what best practice to encourage and second, how to provide that encouragement with the minimum of disruption to the forum activity.

Best Practice

We have identified three sources from which to consider our advice for post-editors in our initial experiments. Given the diversity of language skills in forum users, we cannot always rely on the editor's ability to reference the source material.

1. Post Editing Guidelines For GALE Machine Translation Evaluation[1]
2. Machine Translation Post-Editing Guidelines – by TAUS, in partnership with CNGL[2]
3. Post-Editing Machine Translated Text in A Commercial Setting, Midori Tatsumi, PhD thesis[3]

We have reviewed these bilingual post-editing guidelines/ monolingual post-editing guidelines with access to a reference translation and adapted them to our use case.

Post Editing Guidelines for GALE Machine Translation Evaluation

Make the MT output have the correct meaning, using understandable English, in as few edits as possible.

1. Make the MT output have the same meaning as the reference human translation. No more and no less.
2. Make the MT output be as understandable as the reference.
3. Capture the meaning in as few edits as possible using understandable English. If words/phrases/punctuation in the MT output are completely acceptable, use them (unmodified) rather than substituting something new and different.
4. Punctuation must be understandable, and sentence-like units must have sentence ending punctuation and proper capitalization. Do not insert, delete, or change punctuation merely to follow traditional optional rules about what is "proper."

[1] http://projects.ldc.upenn.edu/gale/Translation/Editors/GALEpostedit_guidelines-3.0.2.pdf
[2] http://www.translationautomation.com/machine-translation-post-editing-guidelines.html
[3] http://doras.dcu.ie/16062/1/SAKURA_final_revised.pdf

Machine Translation Post-Editing Guidelines (TAUS)

1. Aim for semantically correct translation.
2. Ensure that no information has been accidentally added or omitted.
3. Edit any offensive, inappropriate or culturally unacceptable content.
4. Use as much of the raw MT output as possible.
5. Basic rules regarding spelling apply.
6. No need to implement corrections that are of a stylistic nature only.
7. No need to restructure sentences solely to improve the natural flow of the text.

Post-Editing Machine Translated Text in a Commercial Setting (Midori Tatsumi)

The post-edited text needs to be easily understandable by the readers. In order to achieve that goal, the text needs to convey the correct meaning of the source text, and conform to Japanese grammar. However, speed is another important requirement for post-editing processes. Therefore, it is not necessary to spend time aesthetically refining the text; please avoid editing for stylistic sophistication.

A. What needs to be fixed:

1. Non-translatable items, such as command and variable names, that have been translated. Please put it back to English.
2. Inappropriately translated general IT terms.
3. Mistranslation (The meaning of the source text has not been conveyed correctly into translation).
4. Word orders that are inappropriate to the extent that the sentence has become impossible or difficult to comprehend.
5. Comprehensible but extremely unnatural or inappropriate expressions.
6. Inappropriate postpositions and conjugations.

There are no detailed rules available on monolingual post-editing. Fluency and clarity seem to be the main goals of monolingual post-editing based on how post-editors interpret the meaning of the target text (as mentioned in Koponen[4]).

From the guidelines expressed above, we have selected the minimum set of best practice advice to present to our forum membership.

Guidelines for Monolingual Post-Editing

- Try and edit the text by making it more fluent and clearer based on how you interpret its meaning.
- For example, try to rectify word order and spelling when they are inappropriate to the extent that the text has become impossible or difficult to comprehend
- If words, phrases, or punctuation in the text are completely acceptable, try and use them (unmodified) rather than substituting them with something new and different.

Guidelines for Bilingual Post-Editing

- Aim for semantically correct translation.
- Ensure that no information has been accidentally added or omitted.

[4] http://www.molto-project.eu/sites/default/files/molto_20110902_mkoponen.pdf

- If words, phrases, or punctuation in the text are completely acceptable, try to use them (unmodified) rather than substituting them with something new and different.

Presentation within a Forum

Best practice presentation in a natural forum format will be attempted so as to keep to an absolute minimum any disruption of the practitioner's core activity.

Linguistic Information for SMT

Forum feedback is a common practice and is encouraged as it serves to inform following analysis by computational linguists of subsequent rounds of MT adaptation within the ACCEPT project.

Community Dissemination

The seminar material has been distributed to the community predominantly via email (see Figure 1) and also as tips from within the ACCEPT portal (see Figure 2).

The below is a communication from TWB to users on how to use the Post-Editing functionality. This included an email and an accompanying PowerPoint.

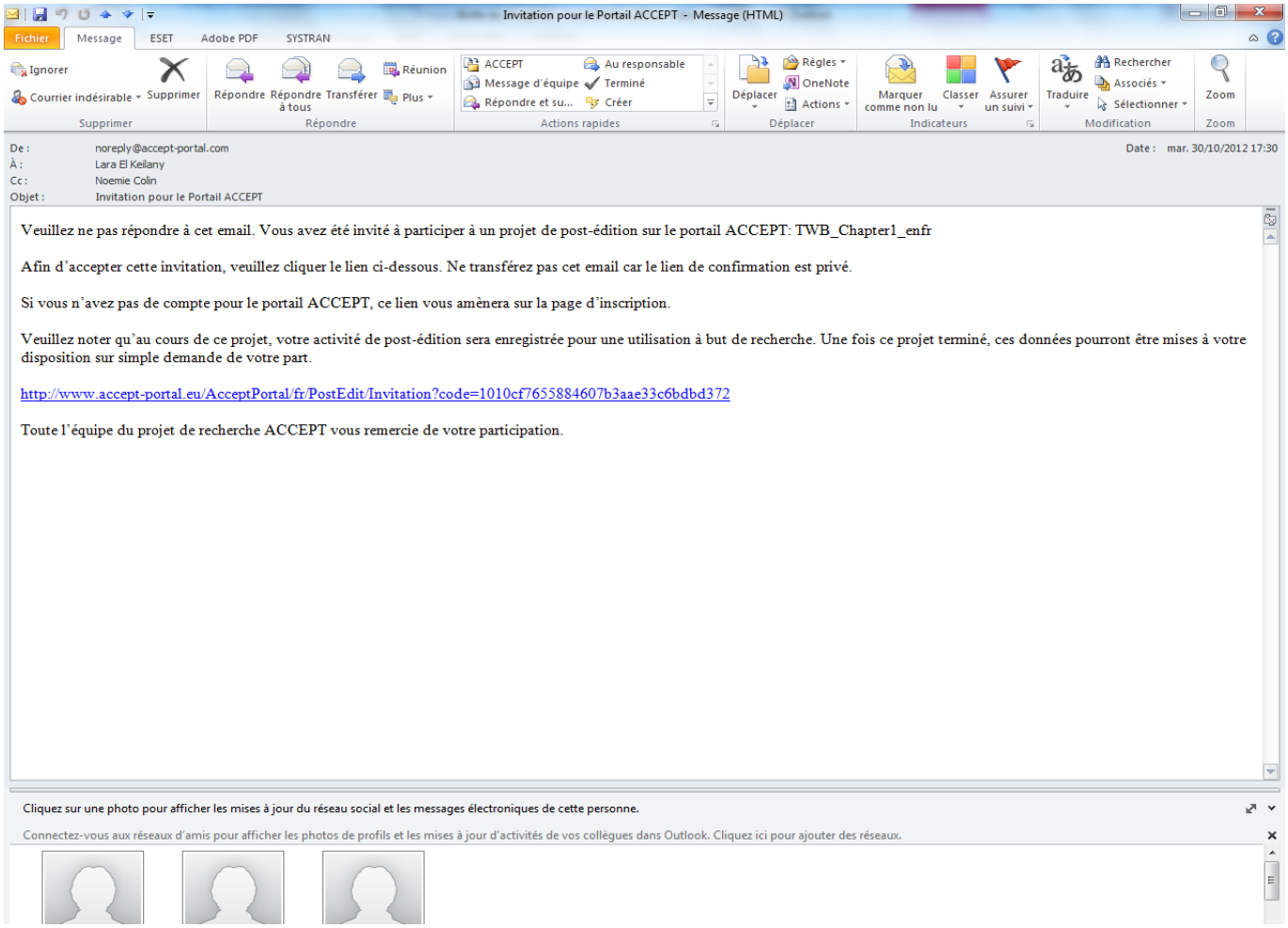

Figure 1: Communication to users on how to use the Post-Editing functionality

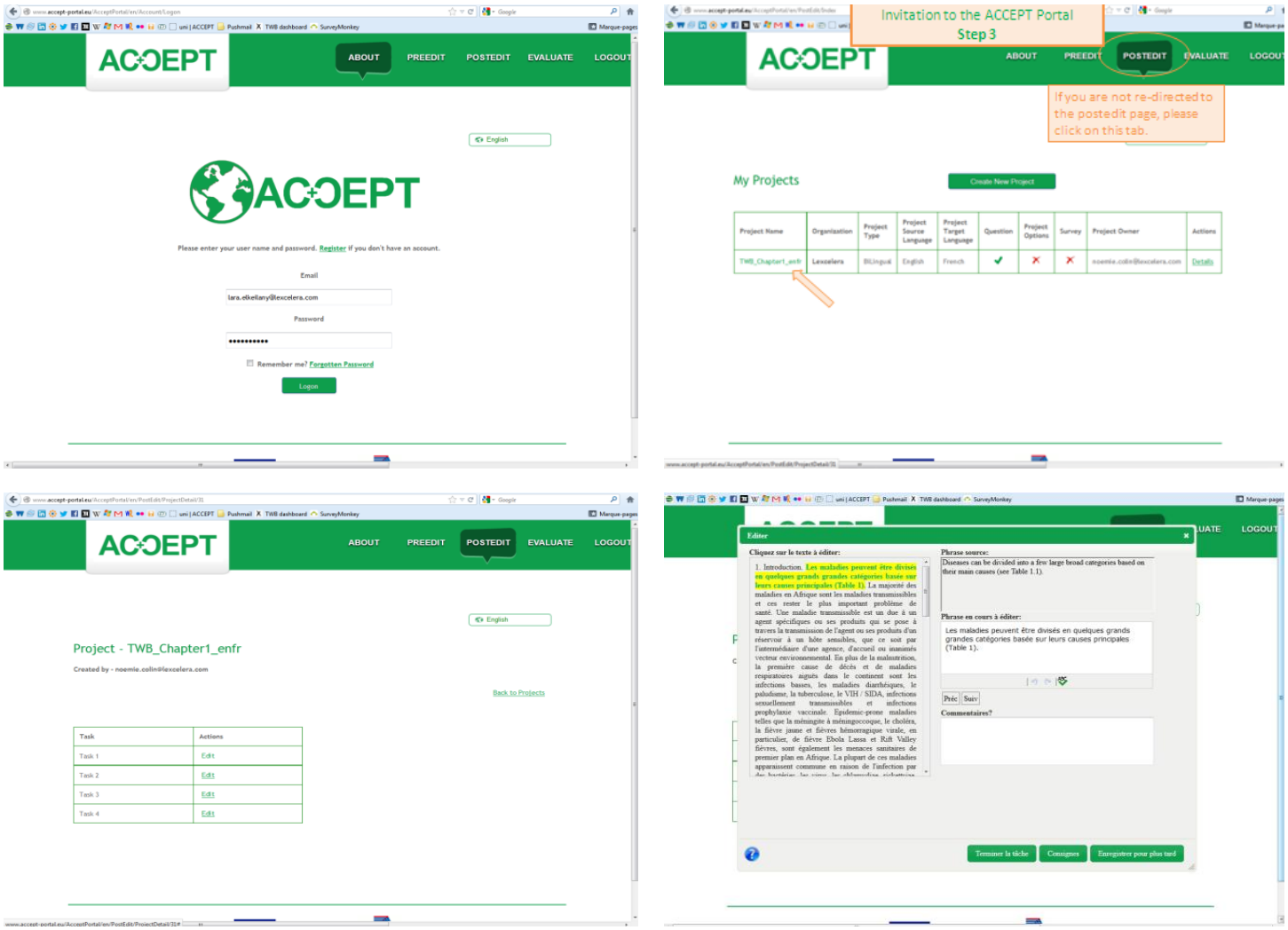

Figure 2: Tips from within the ACCEPT portal

Symantec used a combination of email and the Norton discussion forum to disseminate the seminar material. The focus was on introducing the concept and functionality and took the format of an FAQ.

1. What is MT?

Machine translation systems translate text from one language into another automatically (e.g. like Google Translate). Whoever has used MT before knows that the output is not very good most of the time, as can be seen in the example below. This also shows, however, that even if the German is not correct, you can understand the meaning.

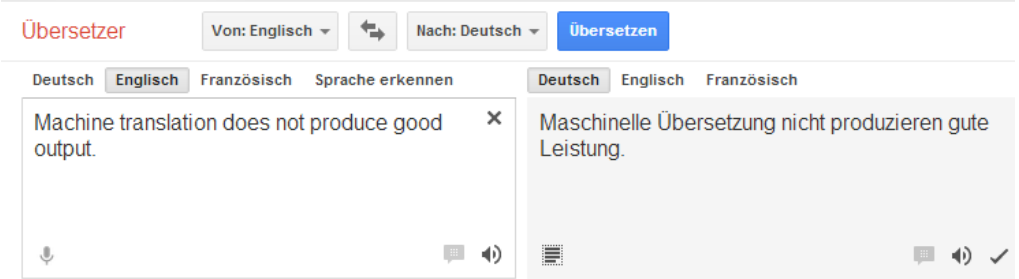

Figure 3: Google Translate

2. What is the ACCEPT project?

This study is conducted by Symantec as part of the ACCEPT project, a European funded project that aims at sharing content across the language barrier more effectively. More specifically, this applies to content that is posted in the English Norton forum by members, which will be made available to users of Norton forums who do not speak English. So people who have the same problem but who do not speak the same language can find the solution to their problem in the translations.

3. Why are we doing this study?

As you can see above and as you can see here *(link to MT part of German forum)*, there is room for machine translation output to be improved. This is why we are conducting this study with you. We want to see whether the editing of/correcting posts improves the sharing process of information across the Norton forums.

4. Your role

Your role will be to edit 14 posts, which have been translated into German. For some of the posts, you will only see the translation, for others, you will see the translation plus alternative translation options and for others again, you will have access to the source.

You will be sent a link to a pre-task questionnaire, which includes some basic questions on you as a member of the forum and which will ask for your email address, so you can be invited to the project. Clicking on the project invitation will bring you to the registration page of the ACCEPT portal and we would kindly ask you to register there. Then you can start editing. During this process you can submit comments. For questions, you can post to the forum here. Once all editing tasks have been completed, there will be a post-task questionnaire on your post-editing experience and motivation.

Guidelines for post-editing:

- Aim for semantically correct translation.
- Ensure that no information has been accidentally added or omitted.
- If words, phrases, or punctuation in the text are completely acceptable, try to use them (unmodified) rather than substituting them with something new and different.

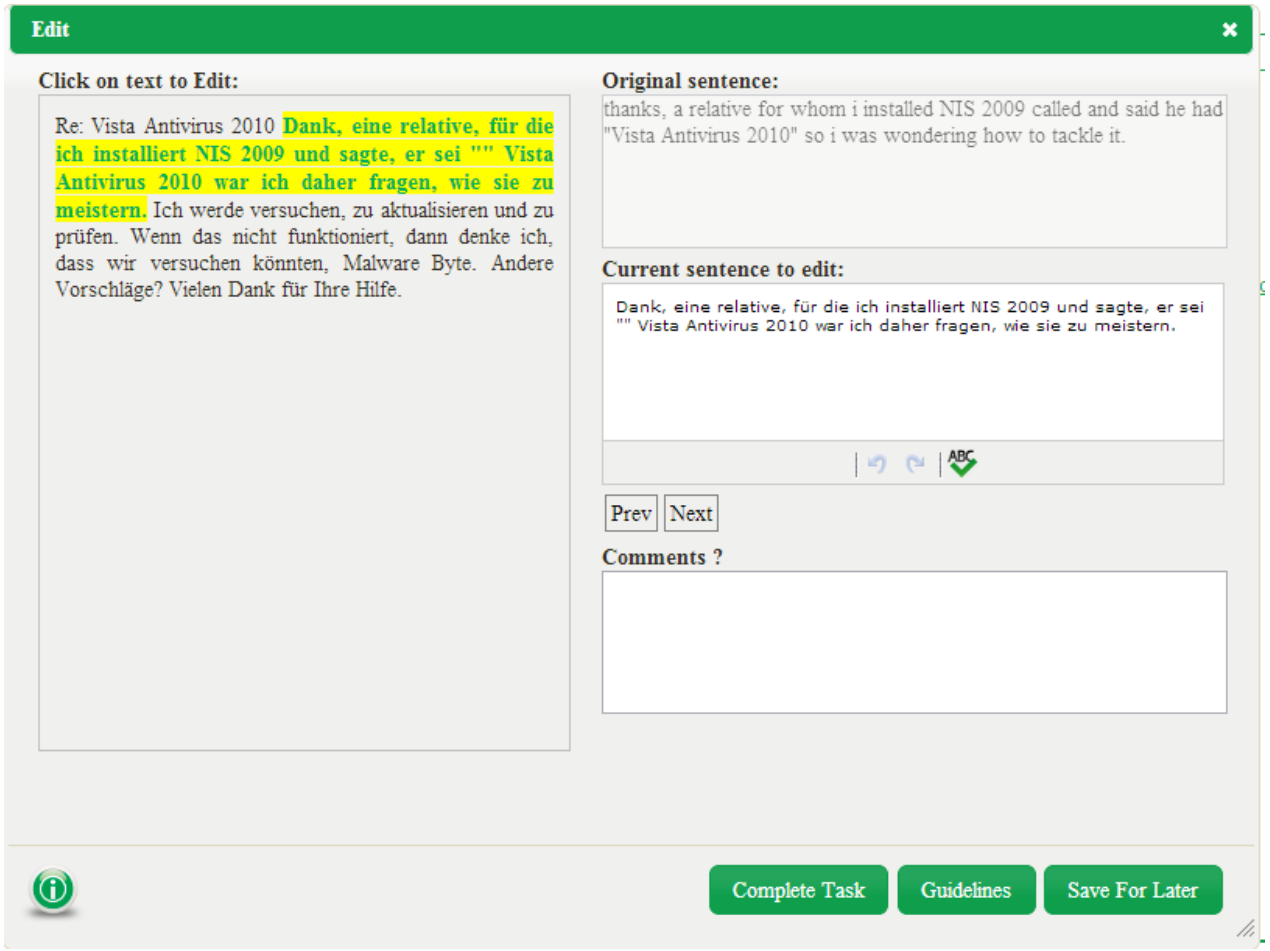

Figure 4: ACCEPT editing window

12 Month Lessons Learned

Although the types of dissemination were well received and enabled those using the functionality to have fewer questions, they were underutilized. For the 6.2.3 we will be increasing the frequency and scope of seminar material to have a larger impact.