

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

Survey of evaluation results – Version 1

Workpackage n° 9

Name: MT Evaluation

Deliverable n° 9.2.2

Name: Survey of evaluation results – Version 1

Due date: 31 December 2013

Submission date: 19 December 2013

Dissemination level: PU

Organisation name of lead contractor for this deliverable: University of Geneva

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

Foreword	3
1 Objectives of the Deliverable	3
2 The Impact of Pre-editing Rules on Translation Quality	3
2.1 Evaluation Methodology	3
2.2 Experimental Setup	3
2.3 Data and Results.....	6
2.3.1 Evaluating the Impact on the Symantec Technical Forum Domain	7
2.3.2 Evaluating the Impact for the TWB Healthcare Domain	9
2.3.3 Automatic vs. Full Checking.....	11
2.3.4 Monolingual vs. Bilingual Evaluation.....	12
2.3.5 3-way vs. 5-way Evaluation	13
2.4 Human Evaluation – Summary of Findings.....	15
2.5 Automatic Evaluation	15
2.6 Task-Based Evaluation.....	16
3 Assessment of User Ratings Reliability.....	17
4 Conclusion	19
References.....	19
Appendix A. Evaluation Tool – Screen Capture	20
Appendix B. Task Guidelines	21
B.1 Comparative Evaluation	21
B.2. Manual Pre-editing.....	23
B.3 Translation.....	24
Appendix C. Post-Task Questionnaire Results.....	26
C.1 Comparative Evaluation	26
C.2 Manual Pre-editing.....	30
Appendix D. Correlation Between Feedback Variables.....	33
Appendix E. Rule Distribution by Category	34

Survey of evaluation results – Version 1

Foreword

As agreed with the Project Officer on 7 May 2013, the original deliverables D9.2.1 (Survey of evaluation results – Version 1) and D9.2.2 (Survey of evaluation results – Version 2) are being merged into the present, common deliverable (D9.2.2).

1 Objectives of the Deliverable

This deliverable provides an account of the (completed) evaluation work carried out in WP9 during month 12 and month 24 in order to assess the technology developed in the ACCEPT project. In this period, the focus of the evaluation has been on quantifying the impact of the ACCEPT pre-editing technology on translation quality (Task 9.1), as the pre-editing rules defined in WP2 constitute one of the major achievements of the project so far. In this deliverable, we describe the evaluation methodology adopted, the evaluation experiments carried out, and the results obtained in this area. In addition, we describe the work focused on a specific evaluation issue, namely, the assessment of user ratings reliability (Task 9.3).

2 The Impact of Pre-editing Rules on Translation Quality

In this section, we provide a survey of the work devoted to evaluating the ACCEPT pre-editing technology in the framework of Task 9.1 *Evaluate the impact of pre-editing rules on SMT* (months 12-18).

2.1 Evaluation Methodology

The evaluation methodology has been initially defined in the deliverable [D 2.1 Definition of pre-editing rules for English and French](#). Accordingly, human judgements are first collected in a contrastive evaluation task involving M.Sc. students in translation, where a 5-point scale is used to rate the translation quality of the raw source text vs. the translation quality of the pre-edited source text: *first clearly better*, *first slightly better*, *about the same*, *second slightly better*, *second clearly better*. Furthermore, automatic metric scores are computed for a subset of the manually-evaluated data for which reference translations have been produced, and their correlation with the manual judgements is reported. This intrinsic evaluation is supplemented by extrinsic evaluation, aimed at assessing the impact of pre-editing on the task of post-editing.

2.2 Experimental Setup

As stated in deliverable D 2.1, there are a large number of conditions in which the evaluation can be performed. The main experimental variables considered in the pre-editing evaluation work are the following:

- domain: {Symantec technical forum, TWB healthcare}
- language pair: {English-French, English-German, French-English}
- MT system: {ACCEPT baseline}
- pre-editing rule set: {automatic, manual, full (both automatic and manual)}

- evaluation type: {bilingual, monolingual}
- manual pre-editing environment: {Word, ACCEPT Portal}
- evaluation scale: {5-point, 3-point}.

The ACCEPT baseline SMT systems are described in the deliverable [D 4.1 Baseline machine translation systems](#). The pre-editing consists of applying, automatically or manually, the correction suggestions proposed by ACCEPT Acrolinx server (accept.acrolinx.com) according to the rules defined in WP2. For the technical forum domain, there are two rule sets defined for English: an automatic rule set, which is typically applied first, and a manual rule set, applied after. For French, there is an additional automatic rule set which is applied at the end and which consists of “silent” rules (not visible to users because they may degrade the quality of the source text). For the healthcare domain, there are two rules sets for English, as above. For French, there is a first manual set, followed by a second set of “silent” rules. (For details, see deliverable [D 2.2 Definition of pre-editing rules for English and French](#)). The pre-editing can take place in Word, if the dedicated Acrolinx plug-in is used, or in the ACCEPT Portal, if the browser-based pre-editing plug-in is used (www.accept-portal.eu).¹ The evaluation type – bilingual or monolingual – refers to the evaluator having or not access to the source text, for reference. Finally, the 5-point evaluation scale is the scale from *first clearly better* to *second clearly better* (as explained above), and the 3-point evaluation scale is the version of this scale where no distinction is made between *clearly* and *slightly*, i.e., *first better, about the same, second better*.

The default experimental set-up is the following: data from all domains and all source languages² are manually pre-edited in Word (by a native speaker) as well as automatically pre-edited, then the raw and the pre-edited source text versions are translated with the baseline system, and, finally, the output translations, if different, are evaluated using the 5-point scale by human judges in a bilingual setting (i.e., the source text is displayed for reference).

The evaluation unit is the whole post, as opposed to the sentence. The main reason for this choice is that we are interested in studying the impact of the application of pre-editing rules in combination, rather than individually (as in WP2). Moreover, it is easier for human judges to evaluate a cohesive text than a sentence taken out of its context. Another advantage is that there is no need for sentence splitting, which is very challenging for user-generated content.

The evaluation is performed using an in-house tool which randomises the order in which the raw and pre-edited source text versions are shown to the user (see screen capture in Appendix A). Additional (dependent) variables used in the experiments are feedback variables and time variables. The feedback variables are the following:

- confidence – how sure the evaluators are that their choice is right: {sure, not so sure}
- difficulty – how difficult it was for them to decide: {easy, difficult}
- importance – how important the difference between the two translations is: {important, very important, not so important}

¹ At the time the evaluation took place, the ACCEPT Portal was not ergonomic enough to allow for the pre-editing of long texts. Therefore, the texts have been pre-edited using the Acrolinx Word plug-in.

² For the TWB healthcare domain, the only language pair considered is French-English.

- low quality – the two translations cannot really be compared because they are incomprehensible: {yes, no}
- conflicts – some parts are better in the first translation, others are better in the second: {yes, no}
- flag – used for marking tricky examples and for adding comments: {yes, no}.

The time variables are:

- the time spent choosing one category in the evaluation scale,
- the time spent providing feedback.

In addition to the main (default) setting described above, evaluation experiments have also been performed in minor conditions, according to specific evaluation scenarios, as described below:

- 1) “Automatic vs. full checking” scenario:
 - o domain: Symantec technical forum
 - o language pairs: English-French, English-German, French-English
 - o MT system: ACCEPT baseline
 - o pre-editing rule sets: automatic, full
 - o evaluation type: bilingual
 - o manual pre-editing environment: Word
 - o evaluation scale: 5-point
- 2) “Monolingual vs. bilingual evaluation” scenario:
 - o domain: Symantec technical forum
 - o language pairs: English-German
 - o MT system: ACCEPT baseline
 - o pre-editing rule sets: full
 - o evaluation type: monolingual, bilingual
 - o manual pre-editing environment: Word
 - o evaluation scale: 5-point
- 3) “3-way vs. 5-way evaluation” scenario:
 - o domain: Symantec technical forum
 - o language pairs: English-German
 - o MT system: ACCEPT baseline
 - o pre-editing rule sets: full
 - o evaluation type: bilingual
 - o manual pre-editing environment: Word
 - o evaluation scale: 3-point, 5-point.

The experiments corresponding to the specific scenarios (1-3) were designed in order to answer the following research questions:

- 1) What is the impact of the automatic pre-editing rules alone, and how does it compare with the impact of the whole rule set, which includes rules requiring manual intervention?

The hypothesis put forward is that the automatic rules might be sufficient for achieving a significant increase in translation quality. The implication is that the ACCEPT pre-editing

technology has a direct, immediate and broadened applicability, regardless of the availability of manual intervention.

- 2) Is monolingual evaluation – i.e., evaluation without access to the source text – feasible? Does it produce comparable results to bilingual annotation?

The hypothesis put forward is that monolingual evaluation is feasible and the results produced without referring to the source are reliable, i.e., comparable with bilingual evaluation results. The implication is that evaluation work can be performed by monolingual speaker participants, who are easier to recruit than bilingual speakers.

- 3) Does the granularity of the evaluation scale have an impact on the evaluation results?

The hypothesis is that when a rougher, 3-point scale is used, the evaluators tend to overuse the *about the same* category; therefore, a separate category *slightly* is necessary for capturing changes that are less important, but still have an impact on translation quality. The implication is the validation of the 5-point scale used in the default experimental set-up.

The experiments relied on the collaboration of numerous participants, who were asked to perform different tasks, from manual pre-editing of data to comparative evaluation and translation. Sample guidelines distributed to the participants for each single task are included in Appendix B. Appendix C presents the results of the post-task surveys conducted to elicit the opinion of participants about the task they performed.

As can be noted, most participants reported that they perceived the experiments as a positive experience; they had no particular difficulties with the domain; and are willing to perform similar tasks in the future. However, they disagreed that the comparative evaluation task was quick and easy and that the amount of data to evaluate was convenient for them. Detailed comments highlighted the fact that it was complicated to evaluate long posts at once, and that the task was time-consuming. The poor quality of the text was a major cause for frustration, as can be seen from the excerpts shown in below.

What made the task very tiring for me was the fact that the source sentences were often already written really badly, because they are forum entries often written by non-native english speakers. This made the automatic translations, which are sometimes already difficult to decipher, even worse. This is mainly why I needed much more time than expected to complete the task.
Sometimes both translations were semi-comprehensible (meaning understandable, but you had to reread them three times to understand because of the weird computer translation).
I don't know how important it is to use forum entries for this experiment, but I think the evaluation would be much easier with correctly written texts.

Figure 1: Participant feedback on the comparative evaluation task ("Detailed comments" excerpts)

2.3 Data and Results

In this section, we report the results obtained in the default evaluation experiments (first for the Symantec technical forum domain, then for the TWB healthcare domain), as well as on the other experiments performed in each of the scenarios presented above. For each experiment, we describe the data used, discuss specific issues encountered, provide statistics on the inter-annotator agreement, present results and interpret them in terms of statistical significance.

2.3.1 Evaluating the Impact on the Symantec Technical Forum Domain

Data. The data in this experiment consists of forum posts actually generated by the Norton Forum Community (<http://community.norton.com/norton>), and made available by our project partner, Symantec. A set of 2000 posts was randomly sampled for each source language, English and French, from an unseen subset of the Symantec data (i.e., a subset which has not been used for development purposes, such as training SMT systems or defining pre-editing rules). To facilitate the processing of the data, the posts were pre-processed by converting <p> tags to newline characters, removing HTML elements, and replacing URLs with placeholders to prevent their automatic translation (e.g., a string like *highlight* being replaced by *souligner* in the example from Figure 2).

Table 1 provides statistics on the dataset considered for each source language. Figure 2 shows a sample forum post in the original format, and Figure 3 the same post after pre-processing.

	Sample size (posts)	Total unseen data (posts)	Average sample post size (words)
English	2000	7064	88.7
French	2000	8393	78.4

Table 1: Symantec technical forum data: statistics

```
Re: restoring a bootable operating drive from an independent recovery point<P>Check these
instructions by Brian</P><P>There is a quirk that it fails the first time.</P><P><A
href="http://community.norton.com/t5/Other-Norton-Products/Network-restore-with-Ghost-15/m-
p/579844/highlight/true#M41167"      target="_blank">http://community.norton.com/t5/Other-
Norton-Products/Network-restore-with-Ghost-15/m-p/579844/highlight/true#M41167</A></P>
```

Figure 2: Sample forum post in the original format

```
Re: restoring a bootable operating drive from an independent recovery point
Check these instructions by Brian
There is a quirk that it fails the first time.<URL>
```

Figure 3: Sample forum post after pre-processing

Inter-annotator agreement. A first portion of the data amounting to 500 posts for each language pair was evaluated by teams of three judges using the methodology presented in Section 2.1 (the remaining posts were evaluated by a single judge; we will refer to these posts as to the second portion of the data). Table 2 displays the inter-annotator statistics between pairs of annotators (Cohen’s k) and between all the three judges (Fleiss’ k), for each language pair, for the first portion of the data.

The agreement is reported both for the original 5-point evaluation scale and for a rougher 3-point scale, in which no distinction is made between the *clearly* and *slightly* categories. As a matter of fact, to report the impact of pre-editing on translation quality, we use the 3-point scale which corresponds to a distinction between positive impact (*second better*), negative impact (*first better*), and no impact (*about the same*).

5-way	Annotators	Agreement statistics (Cohen/Fleiss k)			Agreement (observed)		
French-English	3	0.30			30.4%		
	pairs	0.32	0.35	0.25	50.4%	53.0%	47.4%
English-French	3	0.19			14.0%		
	pairs	0.07	0.11	0.21	29.0%	31.0%	39.8%
English-German	3	0.20			19.4%		
	pairs	0.14	0.27	0.22	34.8%	43.8%	45.2%

a)

3-way	Annotators	Agreement statistics (Cohen/Fleiss k)			Agreement (observed)		
French-English	3	0.43			57.8%		
	pairs	0.43	0.41	0.47	70.0%	70.6%	69.4%
English-French	3	0.20			28.4%		
	pairs	0.17	0.14	0.31	46.2%	43.8%	55.4%
English-German	3	0.38			46.8%		
	pairs	0.33	0.43	0.38	59.2%	65.2%	61.0%

b)

Table 2: Inter-annotator agreement statistics (Symantec technical forum data, 500 posts, full pre-editing): 5-way=annotation using a 5-point scale; 3-way=annotation using collapsed categories (no distinction between *clearly* and *slightly*). Cohen k values are displayed for agreement between pairs of annotators, and Fleiss k values for agreement between all three annotators.

The relatively low values obtained (up to 0.47, i.e., moderate agreement) are indicative of the difficulty and subjectivity of the task. Evaluators' comments highlighted the difficulty of evaluating long, poor quality texts with conflicting changes. The analysis of the feedback variables showed, indeed, a substantial correlation between difficulty and conflicts (see Appendix D). A previous similar experiment showed that a higher inter-annotator agreement can be achieved for the domain considered when the evaluation unit is the sentence ($k = 0.53$, moderate agreement; Gerlach et al., 2013a).

Impact of pre-editing. We use two different ways of computing the impact of pre-editing on translation quality by taking into account the labels chosen by the three annotators. First, we consider as a reference label the label unanimously chosen by the three judges in a team ("unanimous label"). Alternatively, we consider as a reference label the label chosen by at least two judges of a team ("majority label").

Table 3 reports the impact of pre-editing on translation quality according to human judgements, when the majority label is taken into account.

Majority label	French-English	English-French	English-German
better	68.9%	51.5%	56.4%
same	16.3%	21.7%	14.4%
worse	14.8%	26.9%	29.2%
N	472	443	459

Table 3: Impact of pre-editing on translation quality for the technical domain according to human judgements (first portion of data, majority label). N=number of posts to which a majority label could be assigned

Similarly, Table 4 reports the impact of pre-editing when the unanimous label is taken into account.

Unanimous label	French-English	English-French	English-German
better	82.7%	62.0%	65.4%
same	3.8%	12.0%	5.6%
worse	13.5%	26.1%	29.1%
N	289	142	234

Table 4: Impact of pre-editing on translation quality for the technical domain according to human judgements (first portion of data, unanimous label). N=number of posts to which a unanimous label could be assigned

As for the second portion of the data, currently, the French-English and the English-French language pairs have been investigated. The results are based on a single label, as there was only one evaluator for this portion of the data. Table 5 displays the results on the entire test set, when a single label is taken into account.

Label	French-English	English-French
better	53.9%	49.8%
same	30.0%	23.1%
worse	16.1%	27.1%
N	1756	1569

Table 5: Impact of pre-editing on translation quality for the technical domain according to human judgements (all data, unique label). N=number of posts in the dataset whose translation is affected by pre-editing

Statistical significance. A McNemar test was conducted to compare the number of cases in which the translation became better vs. worse due to pre-editing. For all language pairs, the difference is statistically significant, $p < 0.0001$ (when both the majority label and the unanimous label are taken into account, and both portions of the data are investigated).

2.3.2 Evaluating the Impact for the TWB Healthcare Domain

Data. The data in this experiment consist of 100 sentences randomly selected from a collection of documents authored by doctors and made available by the project partner Lexcelera, through the Traducteurs sans Frontières community of translation volunteers working for NGOs. The document collection provided is highly heterogenous. For the purposes of the project, we selected healthcare reports from the Médecins du Monde NGO.

The data presented specific challenges inasmuch as the conversion of the various formats of document into the text format was concerned, but are much better written than forum data. Before

sampling, the data was filtered such that only the sentences that are neither too short nor too long have been retained (length between 100 and 500 characters). Table 6 shows statistics about the data. A sample sentence is displayed in Figure 4.

	Sample size (sentences)	Total data (sentences)	Average sample sentence size (words)
French	100	2511	29.1

Table 6: TWB healthcare data: statistics

Développer un partenariat avec les collègues de santé mentale concernant épilepsie et infirmités motrices cérébrales, et les violences faites aux enfants, pas assez prise en compte dans les programmes MSF.

Figure 4: Sample sentence from the TWB healthcare domain

Inter-annotator agreement. Two annotators evaluated the translations of the original version and of the pre-edited version of sentences in the dataset.³ As in the Symantec experiment, the annotators were advanced MSc students in translation, native speakers of the target language who are proficient in the source language. The inter-annotator agreement statistics are presented in Table 7. As before, we report the statistics both for the original 5-point scale and the version in which the *clearly* and *slightly* categories are collapsed. The values obtained correspond to fair and moderate agreement. They are slightly higher than those obtained for the Symantec domain, reflecting the fact that the text to evaluate is shorter, with less conflicting changes, and possibly better translated.

	Agreement statistics (Cohen k)	Agreement (observed)
5-way	0.39	53.0%
3-way	0.54	70.0%

Table 7: Inter-annotator agreement statistics (TWB healthcare data, 100 sentences): 5-way=annotation using a 5-point scale; 3-way=annotation using collapsed categories (no distinction between *clearly* and *slightly*).

Impact of pre-editing. We report the results obtained in terms of percentage of better translation, same and worse translation due to pre-editing counting only the cases where the two annotators agree. The impact of pre-editing is shown in Table 8.

Label	French-English
better	50.0%
same	24.3%
worse	25.7%
N	70

Table 8: Impact of pre-editing on translation quality for the healthcare domain according to human judgements. N=number of cases on which the two judges agreed

³ There are two pre-editing rule sets defined in WP2 for the TWB domain; see deliverable [D 2.2 Definition of pre-editing rules for English and French \(final version\)](#). The first set, Portal_Set_1_TWB, contains manual pre-editing rules. The second set, Portal_Set_2_TWB, contains automatic pre-editing rules.

Statistical significance. A McNemar test was conducted to compare the number of cases in which the translation became better vs. worse due to pre-editing. The difference is statistically significant, $p < 0.05$.

2.3.3 Automatic vs. Full Checking

Data. In order to compare the impact of automatic pre-editing rules alone with the impact of the full set of pre-editing rules (including rules which require manual intervention), we randomly selected a set of 100 posts from the whole dataset of 2000 posts used in the Symantec scenario, and let the same teams of judges evaluate the additional 100 posts. The new evaluation task took place roughly at the same time as the main evaluation task.

Inter-annotator agreement. The agreement statistics for the automatically pre-edited dataset are shown in Table 9. The values obtained are comparable with those reported for the main experiment, involving fully pre-edited data.

5-way	Annotators	Agreement statistics (Cohen/Fleiss k)			Agreement (observed)		
French-English	3	0.30			26.3%		
	pairs	0.31	0.27	0.39	43.4%	40.8%	55.3%
English-French	3	0.13			11.8%		
	pairs	0.16	0.09	0.19	35.3%	29.4%	36.8%
English-German	3	0.20			17.8%		
	pairs	0.10	0.34	0.18	30.1%	47.9%	38.4%

a)

3-way	Annotators	Agreement statistics (Cohen/Fleiss k)			Agreement (observed)		
French-English	3	0.47			56.6%		
	pairs	0.42	0.42	0.68	61.8%	61.8%	84.2%
English-French	3	0.26			38.2%		
	pairs	0.27	0.27	0.28	54.4%	58.8%	57.4%
English-German	3	0.40			49.3%		
	pairs	0.35	0.49	0.36	58.9%	68.5%	61.6%

b)

Table 9: Inter-annotator agreement statistics (Symantec technical forum data, 100 posts, automatic pre-editing): 5-way= annotation using a 5-point scale; 3-way=annotation using collapsed categories (no distinction between clearly and slightly). Cohen k values are displayed for agreement between pairs of annotators, and Fleiss k values for agreement between all three annotators.

Comparison of monolingual and bilingual evaluation results. We computed Spearman's correlation coefficient between the labels for the automatically pre-edited posts and the labels for the fully pre-edited counterparts. (This was only possible for the language pairs French-English and English-French, for which annotations were available for all 2000 posts). The results show a significant *moderate/strong* correlation between the two label sets ($p < 0.01$), both when the 5-point scale is considered (English-French: Spearman's $\rho = 0.641$; French-English: Spearman's

rho = 0.575), and when its 3-point version is considered (English-French: Spearman’s rho = 0.64; French-English: Spearman’s rho = 0.532).

Impact of pre-editing. In Table 10 we report the impact of automatic pre-editing on translation quality, when the majority label is taken into account (the reference label is the one chosen by at least two judges in a team).

Majority label	French-English	English-French	English-German
better	64.3%	64.1%	54.5%
same	5.7%	12.5%	10.6%
worse	30.0%	23.4%	34.8%
N	70	64	66

Table 10: Impact of automatic pre-editing on translation quality for the technical domain according to human judgements. N=number of post whose translations were affected by pre-editing and to which a majority label could be assigned

Statistical significance. McNemar tests were conducted to compare the number of cases in which the translation became better vs. worse due to automatic pre-editing. For the French-English and English-French language pairs, the difference is statistically significant ($p < 0.01$). For English-German, it is not statistically significant, hence the particular importance of manual pre-editing for this pair of languages.

2.3.4 Monolingual vs. Bilingual Evaluation

Data. In order to test whether monolingual evaluation is feasible and whether the results of monolingual evaluation are comparable with the results of bilingual annotation, we randomly selected 100 posts from the first portion (500 posts) already evaluated in the main Symantec scenario. One of the goals of the ACCEPT project is to focus on monolingual, as opposed to bilingual evaluation, since monolingual subject matter experts are easier to find than bilingual ones. This experiment was designed to test if two competing translations can be reliably compared against each other in the absence of the source text. The experiment was conducted for the English-German language pair (see the “Monolingual vs. bilingual evaluation” scenario in Section 2.2) and took place about 5 months after the main evaluation experiment. Statistics about the data used in the experiment are shown in Table 11.

	Sample size (posts)	Average sample post size (words)
English	100	105.9

Table 11: Data used in the monolingual evaluation experiment

Intra-annotator agreement. The same annotator who evaluated the 500 posts in the main evaluation task re-evaluated the 100 posts in a monolingual setting. We report the intra-annotator agreement statistics in terms of Cohen’s k and observed agreement between two label sets, the initial and the new one. Table 12 displays these statistics for both the original 5-point evaluation scale and for the 3-point scale.

	Agreement statistics (Cohen k)	Agreement (observed)
5-way	0.26	44.3%
3-way	0.41	68.2%

Table 12: Intra-annotator agreement statistics for the monolingual vs. bilingual label sets: 5-way=annotation using a 5-point scale; 3-way=annotation using collapsed categories (no distinction between *clearly* and *slightly*)

When considering the 3-point scale, we found that 68.2% of the data is annotated with the same label; Cohen’s k is 0.41, i.e., there is a moderate agreement between the initial label in the bilingual setting and the new label in the monolingual setting. We interpret these results as indicative of the feasibility of the monolingual evaluation task and of the reliability of its results.

Comparison of monolingual and bilingual evaluation results. We computed Spearman’s correlation coefficient between the two sets of labels (collected in a monolingual vs. bilingual evaluation setting). The results show a significant moderate correlation between the two label sets ($p < 0.01$), both when the 5-point scale is considered (Spearman’s rho = 0.536), and when its 3-point version is considered (Spearman’s rho = 0.490).

Impact of pre-editing. The impact of pre-editing on translation quality, according to human judgements collected in a monolingual setting, is shown in the second column of Table 13. The third column displays the results obtained for the same data when evaluated in a bilingual setting.

Label	Monolingual evaluation	Bilingual evaluation
better	60.2%	61.4%
same	13.6%	11.4%
worse	26.1%	27.3%
N	88	88

Table 13: Impact of pre-editing on translation quality for the technical forum domain according to human judgements collected in a monolingual vs. bilingual evaluation setting. N=number of posts in the dataset whose translation is affected by pre-editing

Statistical significance. A McNemar test was conducted to compare the number of cases in which the translation became better vs. worse due to pre-editing, when the monolingual evaluation results are taken into account. The difference is statistically significant, $p < 0.001$. Similarly, when the bilingual evaluation results for the same dataset are considered, the difference is again significant, $p < 0.001$.

2.3.5 3-way vs. 5-way Evaluation

Data. To assess the effect of the granularity of the evaluation scale on the evaluation results, we performed an experiment in which 100 randomly selected posts from those used in the main Symantec scenario were re-evaluated using a rougher 3-point evaluation scale instead of the initial 5-point scale:

- initial scale: *first clearly better, first slightly better, about the same, second slightly better, second clearly better;*
- new scale: *first better, about the same, second better.*

The experiment was conducted for the English-German language pair (see the “3-way vs. 5-way evaluation” scenario in Section 2.2) and took place about 5 months after the main evaluation experiment. Statistics about the data used in the experiment are shown in Table 14.

	Sample size (posts)	Average sample post size (words)
English	100	107.4

Table 14: Data used in the 3-way evaluation experiment

Intra-annotator agreement. The same annotator who performed the evaluation in the main Symantec scenario re-evaluated the subset of 100 posts in a 3-way evaluation setting. In Table 15 we report the intra-annotator agreement statistics in terms of Cohen’s k and observed agreement between the two label sets, the initial and the new one.

	Agreement statistics (Cohen k)	Agreement (observed)
3-way	0.34	58.6%

Table 15: Intra-annotator agreement statistics for the 3-way vs. 5-way evaluation label sets. The categories in the initial set are collapsed (no distinction between *clearly* and *slightly*)

Comparison of 3-way and 5-way evaluation results. Spearman’s rho correlation coefficient computed between the two label sets, one corresponding to 5-way and the other to the 3-way evaluation, shows that there is significant *moderate* correlation between these label sets (Spearman’s $\rho = 0.462, p < 0.01$).

The confusion matrix summarising the agreement between the two label sets is shown in Table 16. As it can be noted, in 10 cases the evaluator switched from a *slightly* category (*first slightly better* or *second slightly better*) to an *about the same* category. There were a total of 13 *about the same* labels in the 5-way evaluation; when a 3-point scale was used, the number of *about the same* labels went up to 23. This may suggest that indeed, evaluators seem to overuse the *about the same* category when provided with a rougher evaluation scale. However, the difference observed is not statistically significant, according to the McNemar test. This means that the choice of the granularity of the scale does not bear a significant impact on the evaluation results obtained, confirming the finding above.

	first better	about the same	second better
first clearly better	6	4	1
first slightly better	8	4	5
about the same	4	6	3
second slightly better	6	6	19
second clearly better	0	3	12

Table 16: Confusion matrix for the 5-way and the 3-way label sets.

Impact of pre-editing. The impact of pre-editing on translation quality, according to human judgements collected in a 3-way evaluation setting, is shown in the second column of Table 17. The third column displays the results obtained for the same data in a 5-way evaluation setting.

Label	3-way evaluation	5-way evaluation
better	46.0%	52.9%
same	26.4%	14.9%
worse	27.6%	32.2%
N	87	87

Table 17: Impact of pre-editing on translation quality for the technical forum domain according to human judgements collected in a 3-way vs. 5-way evaluation setting. N=number of posts in the dataset whose translation is affected by pre-editing

Statistical significance. A McNemar test was conducted to compare the number of cases in which the translation became better vs. worse due to pre-editing, when the 3-way evaluation results are taken into account. According to the results of this test, the difference is not quite statistically significant ($p = 0.0608$). When the 5-way evaluation results for the same dataset are considered, the difference is statistically significant ($p = 0.0481$).

2.4 Human Evaluation – Summary of Findings

Human evaluation experiments have been performed on both domains considered, namely, the Symantec technical forum domain and TWB healthcare domain. The experiments investigated the impact of pre-editing on translation quality by taking into account relative ratings on a 5-point evaluation scale. The comparative judgements were collected in a bilingual evaluation setting, i.e., with access to the source text. A statistically significant increase in translation quality was found for both domains and for all language pairs considered.

Additional human evaluation experiments were performed for the Symantec technical forum domain in minor conditions (automatic pre-editing only, monolingual evaluation – i.e., evaluation without access to the source text – and evaluation using a 3-point evaluation scale). It was found that automatic pre-editing alone is sufficient for attaining a statistically significant increase in translation quality for the French-English and English-French language pairs, but not for English-German, where manual pre-editing seems to be particularly important. Monolingual evaluation was found feasible and comparable in results to bilingual evaluation. Another finding was that the granularity of the evaluation scale did not have a high impact on the results, the 3-way and 5-way evaluation showing comparable results.

2.5 Automatic Evaluation

The impact of pre-editing rules on translation quality is also quantified by taking into account automatic metric scores. The metrics used are BLEU, GTM, METEOR and TER, selected according to the DOW and reviewed in the deliverable [D 9.1 Analysis of existing metrics and proposal of a task-oriented metric](#).

Metric scores were computed on a subset of the manually-evaluated data, for which reference translations have been produced. This subset consists of 50 forum posts in French, randomly selected among the 2000 posts considered in the main human evaluation experiment, such that they are likely to represent useful posts (according to the work on text classification performed in WP3, a feature indicating useful posts is the length of the posts, if higher than 186 characters; see deliverable [D 3.1 Taxonomy of forum content and rules for automatic classification](#)). This usefulness criterion

was applied in order to better focus the translation effort on those posts deemed to be worth processing.

The selected posts were translated into English by an advanced MSc student in translation. Statistics about the data are shown in Table 18 below.

	Size (words)	Average post size (words)
source (French)	2616	26.16
target (English)	2554	25.54

Table 18: Reference data for automatic evaluation: statistics

The metric scores were computed using the implementation available in the Asiya online tool (http://asiya.lsi.upc.edu/demo/asiya_online.php). For each post, we retrieved the document-level metric score. To evaluate the impact of pre-editing, we compared the scores obtained for the translation of the raw source text with the scores for the translation of the pre-edited version. The Kendall's tau correlation coefficient was used to measure the correlation between the difference in score, on the one hand, and the relative rating of posts as assigned by human judges. The results for each of the metrics considered are summarised in Table 19.

	Kendall's tau
BLEU	0.174
GTM	0.130
METEOR	0.211
TER	0.181

Table 19: Correlation between automatic metric scores and human judgements

The results show non-significant *weak/weak or no* correlation between human judgements and automatic metric scores, which merely confirms known findings in the literature (e.g., Koehn, 2010). The values obtained are in line with those reported in similar studies in the literature (e.g., Specia et al., 2010). They allowed us to identify the best suited metric to our evaluation scenario: the METEOR metric has the highest correlation with human judgements for the particular domain (Symantec technical forum), language pair (French-English) and dataset considered.

2.6 Task-Based Evaluation

The intrinsic evaluation of the ACCEPT pre-editing technology is supplemented by an extrinsic evaluation, concerned with assessing the impact of pre-editing on a particular task, namely, the post-editing of machine translation results.

An experiment was designed in order to compare post-editing productivity for pre-edited text with that for raw source text. The experiment was performed on a dataset from the technical forum domain containing representative sentences sampled from the French Norton forum data provided by the project partner, Symantec. The dataset consists of 684 sentences, from which a subset of 158 sentences was post-edited by three native English speakers. These sentences are selected to include only those that had a positive pre-editing impact on translation quality, according to unanimous judgements collected from three bilingual judges in a comparative evaluation task similar to the ones reported in the previous sections.

Post-editors were asked to modify the translation of the raw source and the translation of the pre-edited source by performing minimal changes such that the final target sentences were grammatical and conveyed the same meaning as the source sentences. Each post-editor processed both translation versions, and the processing order was randomised. The post-editing effort in terms of time and keystrokes was recorded. The sentences for which the raw translation was processed first slightly outnumbered those for which the pre-edited translation was proposed first (89 vs. 69). To balance the dataset with respect to processing order, the sentences in excess were withdrawn.

For the remaining 138 sentences with their two translation counterparts, the average post-editing speed for the three post-editors showed an increase from 27.7 words/min to 51.7 words/min due to pre-editing (the difference is statistically significant). The average post-editing time is basically reduced by half thanks to pre-editing (more precisely, it is multiplied by a factor of 0.53). When taking into account the time spent pre-editing the source, the results show that the combined pre-editing and post-editing time still correspond to an increase in the average processing speed, from 27.7 words/min to 36.8 words/min.

The automatic TER metric scores computed using the post-edited sentence versions as references also reflected an improvement due to pre-editing (20.17 for the translations corresponding to the raw source vs. 10.76 for the ones corresponding to the pre-edited source; note that lower values indicate an improvement).

The results show that pre-editing rules that improve the translation quality also have an impact on the post-editing productivity. The detailed presentation of the experiment and findings can be found in Gerlach et al. (2013b).

3 Assessment of User Ratings Reliability

This section, describes work devoted to the assessment of the reliability of user ratings, corresponding to Task 9.3 (months 18-24).

One of the concerns of evaluation work in the ACCEPT project is whether judgments collected from end users are reliable, that is, whether they correlate significantly with judgements elicited from translators.

In order to verify this correlation, we carried out a study of the data collected in a previous experiment, which dealt with the individual evaluation of pre-editing rules in WP2 (Gerlach et al., 2013a).

The data used in this experiment are a subset of the Symantec technical forum data. They consist of 1313 French sentences, pre-edited then translated into English using the ACCEPT baseline system. For each sentence, two teams of annotators compared the translation of the original version with the translation of the pre-edited version, using the same tool as and same evaluation scale as in the experiments reported in Section 2. The first team was made up of three MSc students in translation, similarly to the above-mentioned experiments. The second team consisted of three Amazon Mechanical Turk workers, selected to request English native speakers with knowledge of French. While the team of translators remained the same for all data, the team of users changed across sentences; a number of 11 users took part to the evaluation experiment in total.

To compare the judgements of translators with those of users, we took into account the majority label for each team (i.e., the label on which at least two out of the three members of a team agreed). A majority label could be assigned to 94.2% of the sentences in the case of translators. In the case of users, the percentage was slightly higher, 94.7%. The percentage of sentences that received a majority label from both users and translators is 89.7%, corresponding to 1178 sentences. For the remaining sentences, there is complete disagreement either in the translator team or in the user team.

Table 20 reports the Cohen’s k agreement statistics between the majority label assigned by users and translators. The same label was chosen by translators and users in 82.3% of the cases; the k value shows substantial agreement.

	Agreement statistics (Cohen k)	Agreement (observed)
3-way	0.63	82.3%

Table 20: Statistics for agreement between translator and user judgements. The categories of the original 5-point scale are collapsed (no distinction between *clearly* and *slightly*).

The Spearman’s correlation coefficient is very high, $\rho = 0.754$ ($N=1178$, $p < 0.01$). There is a significant strong correlation between the labels assigned by users and those assigned by translators. This means that the judgements collected from users in the Amazon Mechanical Turk platform are reliable, which bears important implications on the evaluation work in ACCEPT.

A detailed analysis of translators’ judgement reliability was performed at the rule category level. The agreement statistics and the correlation coefficient were computed by taking into account categories of rules, as opposed to the whole set of rules. (Note that since this experiment was focused on evaluating rules individually, each sentence in the dataset corresponding to a single rule). Table 21 shows the results obtained by rule category. The rule distribution by category is presented in Appendix E.

Rule Category	Percentage in test set	Agreement statistics (Cohen k)	Agreement (observed)	Spearman's rho
clitiques	9.0%	0.70	84.0%	0.813
grammaire (accord)	9.3%	0.80	90.8%	0.850
grammaire (autres)	1.5%	0.81	88.9%	0.937
homophones	20.9%	0.65	83.3%	0.757
informel	25.0%	0.62	82.0%	0.735
ordre	4.4%	0.75	86.5%	0.936
ponctuation	15.4%	0.47	74.0%	0.569
reformulation	12.2%	0.59	81.3%	0.764
tu	2.4%	0.47	82.1%	0.715

Table 21: Statistics for agreement and correlation between translators and user judgements, by category of pre-editing rules. The categories of the original 5-point scale are collapsed (no distinction between *clearly* and *slightly*).

These results indicate which category of rules is more prone to disagreement than others (e.g., punctuation); however, on average, there is a substantial inter-annotator agreement (average Cohen's k : 0.65; average observed agreement: 84%) and a significant strong to very strong correlation (average Spearman's ρ : 0.786) between translator judgements and user judgements.

4 Conclusion

The main focus of the evaluation work so far has been on the pre-editing component of the ACCEPT technology, which constitutes one of the main achievements of the project. Intrinsic evaluation has taken into account human judgements and automatic metric scores, whereas extrinsic evaluation has investigated the impact of pre-editing on the task of post-editing. The results of human evaluation show significant, consistent improvement of translation quality due to pre-editing of the source text across the domains and language pairs considered in the project. Automatic evaluation scores do not reflect, however, this improvement. For the metrics investigated, there is *weak/weak or no* correlation between human judgements and metric scores, which merely confirms known findings in the literature and emphasizes, once again, the importance of human evaluation. The improvement in translation quality is accompanied by an improvement in post-editing productivity, our experimental results showing that the time spent post-editing is reduced by half.

In addition to work on pre-editing evaluation, we also reported on work devoted to a specific aspect which is of central importance in our project, namely, the assessment of end user ratings reliability. The results obtained for the pre-editing evaluation scenario – in which the ratings of Amazon Mechanical Turk workers evaluating the relative quality of translations of raw vs. pre-edited sentence versions are compared against those of translators – showed a substantial agreement and a very strong correlation between user and translator judgements. This bears important implications on the ACCEPT project, in which many evaluation experiments rely on user participation.

References

- Gerlach, Johanna, Victoria Porro, Pierrette Bouillon, Sabine Lehmann:
La prédiction avec des règles peu coûteuses, utile pour la TA statistique des forums ?
In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*,
Sables d'Olonne, France, June 2013a.
- Gerlach, Johanna, Victoria Porro, Pierrette Bouillon, Sabine Lehmann:
Combining pre-editing and post-editing to improve SMT of user-generated content
In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, Nice,
France, September 2013b.
- Koehn, Philipp:
Statistical Machine Translation.
Cambridge University Press, 2010.
- Specia, Lucia , Raj Dhwaj, Marco Turchi:
Machine translation evaluation versus quality estimation.
Machine Translation 24(1): 39-50, 2010.

Appendix A. Evaluation Tool – Screen Capture

ACCEPT Comparative Evaluation

pre-edited fr-en-batch.evaluation Browse...

User ID:

Begin evaluation of next Set

Example 3 of 20 in set

Get file stats

Stats: 88 sets in evaluation file, 88 sets to evaluate

Evaluation file

Look at the original sentence and two possible translations (all differences are marked in blue). Choose which of the translations is better.

Original

PB scanner Samsung cxl 3185

Bonjour,

Je viens d'installer un HP touchsmart 1020 sous seven home 64 bits. et y ai installé une imprimante multifonction samsung cxl 3185 qui fonctionne très bien sous paint pour scanner.

Mais si j'utilise le logiciel Scanmgr (présent sous c:\windows\twain32\samsung) j'ai un plantage systématique au lancement

Le PC étant neuf sans logiciel superflu, est il possible que norton internet security soit la cause du problème ?

First translation

Problem scanner Samsung 3185 key

Hi, I just installed a HP TouchSmart 1020 under seven home 64 bitset have installed a Samsung multifunction printer 3185 key that works very well under many to scan. If I use the software Scanmgr (now under c: Windowstwain32samsung) I have a systematic plantage to launch. As the PC is nine without software superfluous, is it possible that Norton internet security is the cause of the problem?

PS I added the file scanmgr.exe in the software allowed in the firewall of Norton. I have also turned off the firewall and antivirus for 15 minutes. but nothing happens.

First clearly better
 First slightly better
 About the same
 Second slightly better
 Second clearly better

Second translation

Pb scanner Samsung cxl 3185

Hi, I just installed a HP TouchSmart 1020 under seven home 64-bit, and have installed a samsung multifunction printer cxl 3185 which works very well under paint to scan. But if I use the software Scanmgr (now under c:\ windows\ twain32\ samsung) I have a systematic plantage to launch the PC being nine without software superfluous, is it possible that norton internet security is the cause of the problem?

PS I added the file scanmgr.exe in the software allowed in the firewall fire of norton. I also have disabled the firewall and antivirus for 15 min but nothing happens.

About the same
 Second slightly better
 Second clearly better

I am sure not so sure that my choice is right.

It was relatively easy difficult to decide.

The difference is not so important important very important

I cannot really compare the two sentences since they are incomprehensible
 Some parts are better in the first sentence; others are better in the second sentence
 This example is tricky; I'd like to add a comment

Previous <

> Next

Save set

Table A.1: Screen capture of the tool used for comparative evaluation

Appendix B. Task Guidelines

B.1 Comparative Evaluation

Manual Evaluation of Automatic Translations for the ACCEPT Project

Guidelines

ACCEPT (<http://www.accept-project.eu>) is a research project devoted to improving machine translation technologies for community content. In particular, it aims to improve the quality of posts from specific Internet forums, such as Symantec's Norton Users Discussion Forum (<http://community.norton.com/>).

To this end, the ACCEPT team has created pre-editing rules for English and French to improve the quality of the source text. We are currently investigating the impact of these rules on the translation quality. We pre-edited a large number of forum posts and we wish to compare the translation obtained for the original posts against the translation obtained for the pre-edited posts. We would greatly appreciate your collaboration on the following task:

Evaluate 25 test sets of 20 translation pairs each

Description - Preparation:

- Download the attached evaluation file (whose name contains "batch" and ends with ".evaluation"). Each evaluation file contains a number of test sets. The exact number may vary.
- Download the executable file ("ComparativeEvaluation.exe"). This is the evaluation tool. Run it by double-clicking on it.
- Enter your name next to "User ID". Open the evaluation file by clicking on "Browse". Click on "Get file statistics" to see the number of test sets to evaluate.
- Start with the first set by clicking on "Begin evaluation of next Set". The interface shows the first example (translation pair) to evaluate: the original forum post, and two competing translations.
- To see the remaining examples, click on "Next". After evaluating all 20 examples, you can save the results by clicking on "Save set". A file identified with your "User ID" will be created.
- Make sure to complete a test set and save it before quitting work. Results are saved only when you click on "Save set".

Your Task:

- Decide which of the two translations is better. The text in blue will help you spot the differences, but remember you are comparing the whole translations. Refer to the original text to make sure you made the right choice.
- When deciding, keep in mind that you will select the translation you would prefer to post-edit.
- After making your choice, fill in the questionnaire on the right side. Please select which of the statements apply in your case.
 - The first three statements are **mandatory**: You have to make a choice.

- The last three statements are **optional**: If they apply in your case, check the box, otherwise leave it empty. Note: You are not requested to enter a comment, but you can do so if you want to flag an example so that we can have a look at it later.

Results:

Please send the results file (whose name end with ".results") by e-mail to Violeta.Seretan@unige.ch by **31 July 2013**. Extensions can be negotiated, if needed.

Payment:

Each test set completed (20 translation pairs) will be paid **5 CHF**.

Contact:

If you need help or have questions or comments, please contact Violeta.Seretan@unige.ch.

Many thanks for your contribution!

B.2. Manual Pre-editing

Manual Pre-editing of User Posts for the ACCEPT project

Guidelines

ACCEPT (www.accept-project.eu) is a research project devoted to improving machine translation technologies for community content. In particular, it aims to improve the quality of posts from specific Internet forums, such as the Norton Users Discussion Forum (community.norton.com).

To this end, the ACCEPT team has created pre-editing rules for English and French, which can be tested on the ACCEPT Portal (www.accept-portal.eu). We are currently investigating the impact of these rules on the translation quality, and plan to pre-edit a large number of posts in order to compare the translation obtained for the original text against the translation obtained for the pre-edited text. We would greatly appreciate your collaboration on the following task:

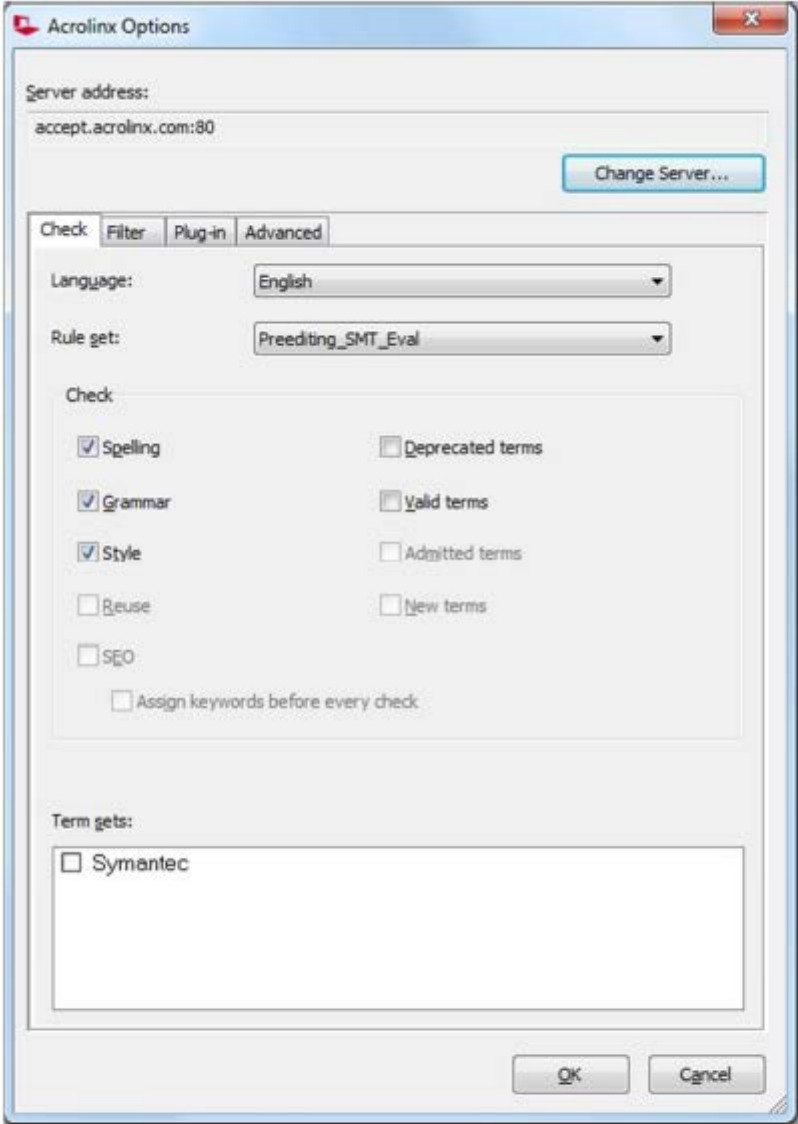
Pre-edit 100 text files of about 1800 words each

Description:

- Download the attached file archive and extract its content. Each file contains 20 posts. There is no relation between these posts; they are selected randomly from our datasets.
- Download the acrolinx plug-in for Word and install it. Open Word, go to the Review section then configure the plug-in option as shown in the attached figure (Appendix 1).
- For each file,
 - Open it in Word, go to the Review section, check that the options are correct, then click on Check. Click on Continue.
 - Identify the highlighted words, which indicate that a rule has been triggered because an error has been detected by the plug-in. Right-click to see the change or the suggestion proposed by the rule.
 - Edit the file according to the change or suggestion proposed, if it improves the text quality.
 - You can use the left and right buttons "Select the previous/next flag" to navigate from error to error.
 - After changing a sentence, for instance by splitting it into shorter sentences, re-check that sentence to see if there are remaining errors. You can ignore an error if no correction is possible.
 - Save the edited file as a text file in UTF-format.
- Archive all the pre-edited files and send the archive by mail to Violeta.Seretan@unige.ch.

Many thanks for your contribution!

Appendix 1. Acrolinx Plug-in Options for English



B.3 Translation

Translation of User Posts for the ACCEPT project

Guidelines

ACCEPT (www.accept-project.eu) is a research project devoted to improving machine translation technologies for community content. In particular, it aims to improve the quality of posts from specific Internet forums, such as the Norton Users Discussion Forum (community.norton.com).

In order to automatically evaluate the technology created for improving the source text quality, we need reference translations for a number of forum posts. We would greatly appreciate your collaboration on the following task:

Translate 50 forum posts of about 50 words each from French to English

Description:

- Download the attached file archive and extract its content. Each file contains one post of about 50 words.
- Create a new folder to store the translations.
- For each file,
 - Open it in Word, translate it, then remove the source text and save the translation in the new folder, using the file name of the original file.
 - Please keep intact the formatting of the source text (line splitting).
 - The source text might contain errors, abbreviations, jargon etc. When translating, feel free to make corrections so that the translated text is as understandable as possible. Example: A sentence like (1) below is to be understood as in (2), therefore in your translation you would use for instance the word *problem* in English rather than the abbreviation *pb*.
 - (1) *Cela aurait il une influence sur mon pb*
 - (2) *Cela aurait-il une influence sur mon problème ?*

Results:

- Please archive the new folder containing the translations and send it by e-mail to Violeta.Seretan@unige.ch no later than **15 August 2013**.

Payment:

The file archive contains 2616 words. The translation will be paid in total 470.88 CHF (corresponding to 0.18 CHF/ word).

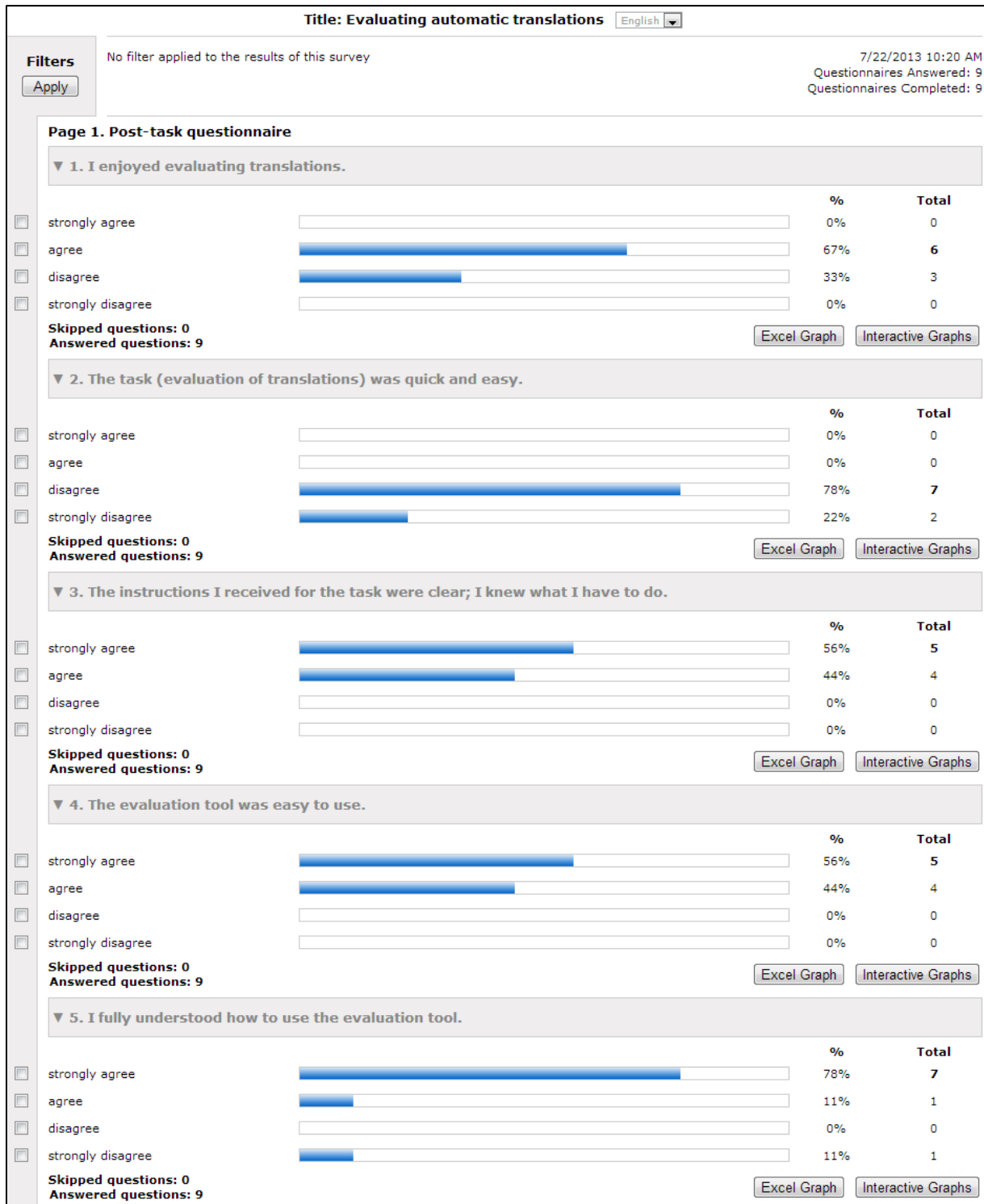
Contact:

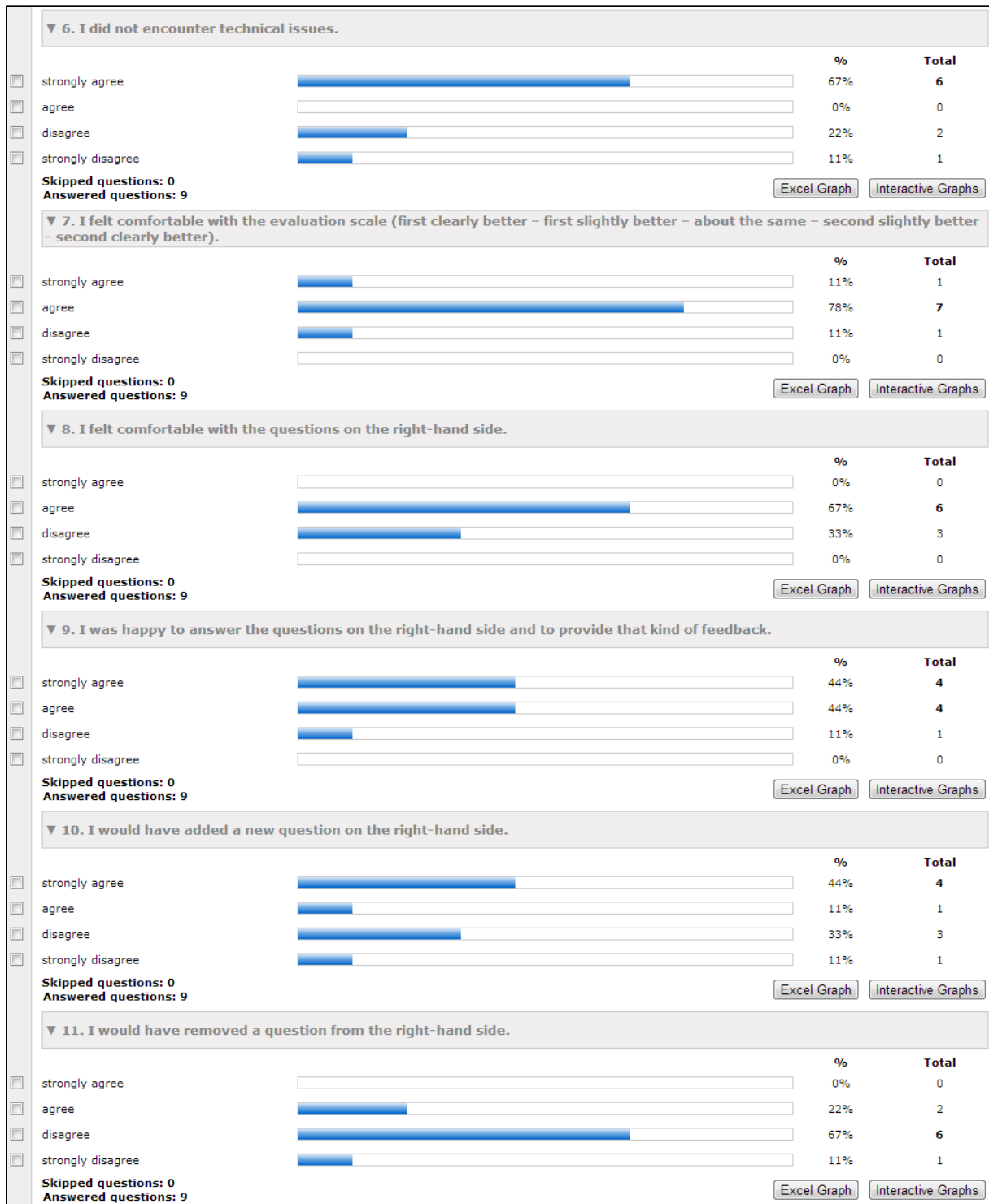
If you need help or have questions or comments, please contact Violeta.Seretan@unige.ch.

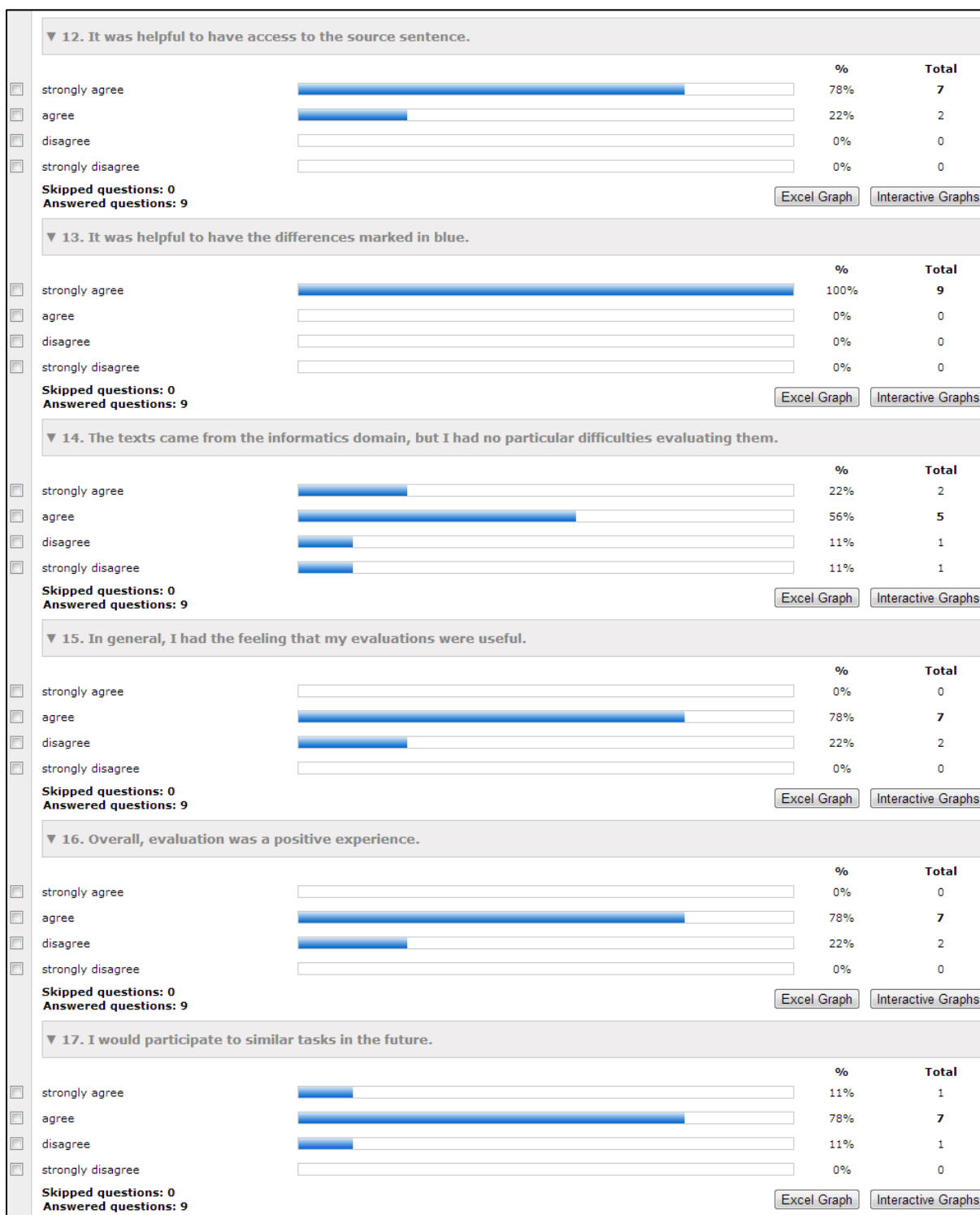
Many thanks for your contribution!

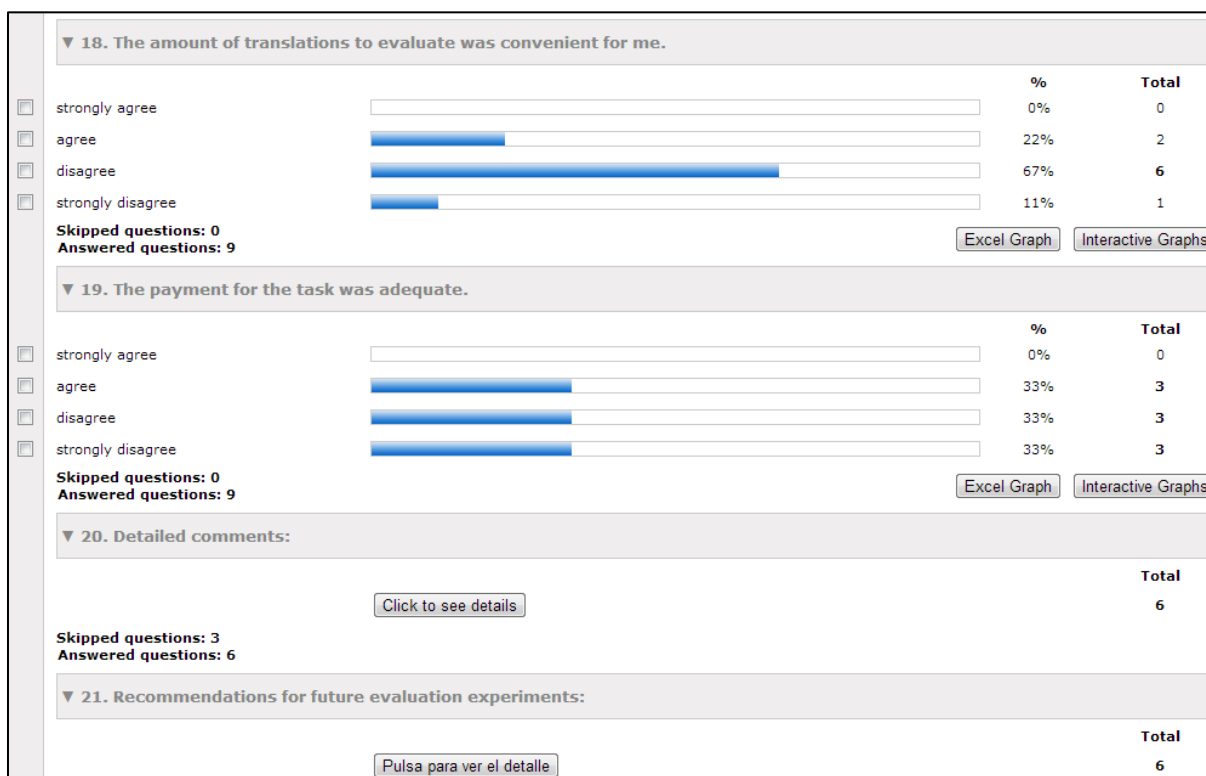
Appendix C. Post-Task Questionnaire Results

C.1 Comparative Evaluation









C.2 Manual Pre-editing

Manual Pre-editing of User Posts

Post-task questionnaire

You recently participated to a task on using a tool for editing forum posts. We would like to know your opinion on this task and the particular editing rules proposed by the tool, so that we can get an idea about how useful these rules are and how you perceived the editing experience.

Please answer the questions below with one of the following answers:

strongly agree – agree – disagree – strongly disagree

For each question, please also enter a comment to explain your choice, if you wish. We would appreciate if you could send the completed document by e-mail to Violeta.Seretan@unige.ch within one week from receiving it.

Thank you again for your collaboration!

1. I enjoyed editing (or editing was fun).

Answer: *agree*

Comment: Since I studied literature and translation, I like when a text is well written, even if I write for strangers on a forum.

2. Editing was quick and easy.

Answer: *strongly agree*

Comment: It was really quick and easy to edit the texts I was given.

3. I did not encounter technical issues.

Answer: *strongly agree*

Comment: It was simple and easy to install all the necessary components to start editing.

4. The instructions I received for the task were clear.

Answer: *strongly agree*

Comment: The instructions were clear and Ms. Seretan always took the time to answer my questions.

5. I understood the purpose of editing the text.

Answer: *agree*

Comment: I understood that the purpose of editing was to turn a text written in chatspeak in a well written text, or at least more understandable by everyone.

6. I understood the suggestions proposed by system.

Answer: *agree*

Comment: -

7. Most of the times, the suggestions proposed were correct.

Answer: *agree*

Comment: Sometimes, the suggestions were not correct but I could understand the reason why the machine misunderstood and gave a wrong suggestion.

8. Most of the times, I knew how to edit the text in order to follow the suggestions.

Answer: *agree*

Comment: -

9. I understood the description of rules (for instance, "subject verb agreement").

Answer: *strongly agree*

Comment: It was easy for me to understand the description of the rules as I have a literature and language background.

10. There are rules where I did not understand what the issue was or what I was supposed to do.

Answer: *disagree*

Comment: -

11. I checked the documentation of a rule when I needed more information.

Answer: *agree*

Comment: It happened for some rules and to edit the text in the most correct way.

12. Most of the rules are useful for correcting the text.

Answer: *agree*

Comment: -

13. There are rules that are not useful for correcting the text.

Answer: *agree*

Comment: There are rules that relate to capital letters (for example *Microsoft* instead of *microsoft*), but I'm not sure it is as important as other types of mistakes.

14. There are rules that I prefer (for instance, "subject verb agreement").

Answer: *agree*

Comment: There are rules that I prefer because they are easier to apply, and the edition is faster with these rules.

15. There are rules that I do not like and I wish they were excluded from the system.

Answer: *strongly disagree*

Comment: -

16. There are new rules I could suggest.

Answer: *disagree*

Comment: I can't think of a rule I could suggest right now.

17. The texts came from the informatics domain, but I had no particular difficulties understanding what I was editing.

Answer: *agree*

Comment: I agree, but sometimes it wasn't always clear what the discussion was about, especially when the person who writes uses lots of abbreviations to talk about some software, for example.

18. In general, I had the feeling that my edits were useful.

Answer: *strongly agree*

Comment: I think my edits were useful since the text was much better in term of linguistic after the editing process.

19. Overall, editing was a positive experience.

Answer: *agree*

Comment: It is a positive experience for someone who likes to write correct sentences.

20. I would participate to similar tasks in the future.

Answer: *agree*

Comment: If I have the time to do it, I would like to participate in the future.

21. The amount of text to edit was convenient for me.

Answer: *agree*

Comment: -

22. The payment received was adequate.

Answer: *agree*

Comment: -

23. I have recommendations for future pre-editing tasks.

Comment: I can't think of a recommendation since all the process went really well as far as I am concerned.

Additional information

Your level (e.g., Master student, 2nd year): Master student in translation

Your expertise in editing: I had basic knowledge but I was a beginner when I edited those texts.

Date: December, 16th, 2013.

Appendix D. Correlation Between Feedback Variables

		Correlations							
		label	confidence	difficulty	importance	toobad	conflicts	flag	time
label	Correlation Coefficient	1.000	.119**	-.186**	.250**	-.192**	-.193**	-.019	-.019
	Sig. (2-tailed)	.	.000	.000	.000	.000	.000	.420	.420
	N	1756	1756	1756	1756	1756	1756	1756	1756
confidence	Correlation Coefficient	.119**	1.000	-.437**	-.064**	-.163**	-.326**	-.256**	-.256**
	Sig. (2-tailed)	.000	.	.000	.007	.000	.000	.000	.000
	N	1756	1756	1756	1756	1756	1756	1756	1756
difficulty	Correlation Coefficient	-.186**	-.437**	1.000	.148**	.244**	.623**	.402**	.402**
	Sig. (2-tailed)	.000	.000	.	.000	.000	.000	.000	.000
	N	1756	1756	1756	1756	1756	1756	1756	1756
importance	Correlation Coefficient	.250**	-.064**	.148**	1.000	.026	.160**	.160**	.160**
	Sig. (2-tailed)	.000	.007	.000	.	.275	.000	.000	.000
	N	1756	1756	1756	1756	1756	1756	1756	1756
toobad	Correlation Coefficient	-.192**	-.163**	.244**	.026	1.000	.160**	.159**	.159**
	Sig. (2-tailed)	.000	.000	.000	.275	.	.000	.000	.000
	N	1756	1756	1756	1756	1756	1756	1756	1756
conflicts	Correlation Coefficient	-.193**	-.326**	.623**	.160**	.160**	1.000	.263**	.263**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.	.000	.000
	N	1756	1756	1756	1756	1756	1756	1756	1756
flag	Correlation Coefficient	-.019	-.256**	.402**	.160**	.159**	.263**	1.000	1.000**
	Sig. (2-tailed)	.420	.000	.000	.000	.000	.000	.	.
	N	1756	1756	1756	1756	1756	1756	1756	1756
time	Correlation Coefficient	-.019	-.256**	.402**	.160**	.159**	.263**	1.000**	1.000
	Sig. (2-tailed)	.420	.000	.000	.000	.000	.000	.	.
	N	1756	1756	1756	1756	1756	1756	1756	1756

** . Correlation is significant at the 0.01 level (2-tailed).

Table D.1: Spearman's rho correlation between feedback variables

Appendix E. Rule Distribution by Category

Rule set	Rule name	Category
3	autoSuggest_utilisezCa	clitiques
3	autoSuggest_utilisezCeuxCi	clitiques
3	évitez_me_m_a	clitiques
2	accord_phrase_nominale	grammaire (accord)
2	accord_sujet_verbe	grammaire (accord)
2	forme_verbale_incorrecte	grammaire (accord)
1	confusion_futur_conditionnel	grammaire (autres)
1	évitez_conditionnel	grammaire (autres)
1	mettez_impératif	grammaire (autres)
2	utilisezSubjonctif	grammaire (autres)
1	a_vs_à	homophones
1	ça_vs_sa	homophones
1	ce_vs_se	homophones
1	ci_vs_si	homophones
1	des_vs_dès	homophones
1	du_vs_dû	homophones
1	expression_incorrecte	homophones
1	la_vs_là	homophones
1	ma_vs_m_a	homophones
1	ou_vs_où	homophones
1	qu_elle_vs_quelle	homophones
1	soit_vs_sois_vs_soi	homophones
1	sur_vs_sûr	homophones
1	tes_vs_t_es	homophones
1	tous_vs_tout	homophones
2	homophones_verbe_nom	homophones
2	séquence_incorrecte_de_mots	homophones
1	erreur_de_majuscule	ignore
2	évitez_est_ce_que	informel
2	évitez_le_langage_familier	informel
2	évitez_le_participe_present	informel
2	évitez_les_anglicismes	informel
2	évitez_les_phrases_clivées	informel
2	évitez_les_questions_directes	informel
2	évitezAbrevForum	informel
2	merci_de_tenir_ac	informel
2	négation_incomplète	informel
3	autoSuggest_tout	ordre
3	évitez_jamais_après_verbe	ordre
3	évitez_rien_avant_infinitif	ordre
1	ajoutez_un_blanc	ponctuation
1	ajoutez_un_trait_d_union	ponctuation
1	ajoutez_une_virgule	ponctuation
1	élidez_ce_mot	ponctuation
1	espaces_autour_ponctuation	ponctuation
1	évitez_ponctuation	ponctuation

1	ponctuation double	ponctuation
2	évitez_le_pluriel_entre_parenthèses	ponctuation
2	fin_de_phrase_sans_ponctuation	ponctuation
2	ne_pas_élider	ponctuation
2	wordDotWord	ponctuation
3	ajoutez_dois_je	reformulation
3	autoSuggest_abreviationIncorrecte	reformulation
3	autoSuggest_avoir_bea	reformulation
3	autoSuggest_evitezMerciDe	reformulation
3	autoSuggest_formules_politesse	reformulation
3	autoSuggest_il_faut_que	reformulation
3	autoSuggest_langage_familier	reformulation
3	autoSuggest_ne_manquez_pas	reformulation
3	autoSuggest_utilisez_seulement	reformulation
3	evitez_verbe_plus_rien	reformulation
3	evitez_tu	tu

Table E.1: Distribution of French pre-editing rules by category. Rule set codes are used to identify the specific pre-editing rule set to which a rule belongs: 1 = Portal_Set_1 (automatic rules), 2 = Portal_Set_2 (manual rules), 3 = Portal_Set_3 (silent automatic rules)