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 2

Workpackage n° 9	Name: MT Evaluation
Deliverable n° 9.2.4	Name: Survey of Evaluation Results – Version 2
Due date: 31 December 2014	Submission date: 19 December 2014

Dissemination level: PU

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

Author(s): Violeta Seretan, Robert Grabowski

Internal reviewer(s): Johann Roturier

Proofreading: Manny Rayner

Copyediting: Violeta Seretan

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

List of Tables

Foreword

As agreed with the Project Officer on 7 May 2013, the original deliverables D9.2.3: *Survey of evaluation results – Version 3* and D9.2.4: *Survey of evaluation results – Version 4* are merged into the present, common deliverable (D9.2.4).

1. Objectives of the Deliverable

This deliverable describes the evaluation work carried out in WP9 during the M25-36 period in order to evaluate the pre-editing and post-editing rules developed in the ACCEPT project. In this period, the focus of the evaluation has been on the impact of post-editing rules on translation quality (Task 9.2), the comparison of the performance of pre-editing rules across MT paradigms (RBMT vs SMT, Tasks 9.4-5), and the analysis of user interaction with pre-editing rules (Task 9.6). Further work has been carried out to extend the evaluation of pre-editing rules, despite the relevant task being closed (Task 9.1). In addition, the results obtained in this period add new insights on the issue of user ratings' reliability and correlation with expert ratings (Task 9.3, closed). Summing up, the evaluation work carried out in the final year of the project addressed all WP9 tasks in the DOW.

This document is organised as follows. We first focus on pre-editing rules and report the new evaluation experiments, which extend the previous work presented in Deliverable D9.2.2 (Section 2). In Section 3, we describe the experiments carried out to assess the performance of pre-editing rules across MT paradigms. Next, we focus on post-editing rules and report on the experiments designed to evaluate this specific component of the ACCEPT technology (Section 4). We describe results on user interaction with the ACCEPT pre-editing rules in Section 5. Section 6 concludes the document.

2. Performance of ACCEPT Pre-editing Rules: New Results

In this section, we describe the new evaluation experiments carried out during the reporting period in three areas. The first area concerns the impact of ACCEPT pre-editing rules on the quality of SMT output for the English-German language pair (Section 2.1). The second area is related to the correlation between automatic and human comparative evaluation results (Section 2.2). Finally, the third area refers to the impact of pre-editing rules on data from a domain different to the one targeted by the project (Section 2.3).

2.1 Impact of Pre-editing on Translation Quality: New Results for English-German

English-German is one of the three language pairs considered in the ACCEPT project, the others being English to French and French to English. Pre-editing rules have been defined in the ACCEPT project using the Acrolinx technology for both source languages considered in the project, English and French (see Deliverable D2.2).[1] The evaluation of pre-editing rules has been described in a number of publications, e.g., Roturier et al. (2012), Gerlach et al. (2013), Bouillon et al. (2014), Seretan et al. (2014), as well as in the project deliverables D2.2 and D9.2.2 (a detailed study of the impact of pre-editing on translation quality is presented in Gerlach, 2015).

Deliverable D9.2.2, in particular, details the large-scale evaluation campaign conducted in Year 2 of the project in order to assess the impact of pre-editing rules on translation quality, taking into account both human judgements and automatic metric scores. The human evaluation experiment involved advanced Master's students in translation at the University of Geneva. The experiment was performed on two large datasets, each consisting of 2000 documents in English and French,

[1] Note that, although the rules are implemented using the Acrolinx formalism and they are made available using the Acrolinx technology, their definition is system-agnostic.

respectively. Each document corresponded to a forum post from a technical user forum related to Symantec's Norton security products (henceforth, the Symantec domain).[2] Each dataset was split into two parts: a first portion of 500 posts, which was evaluated by three annotators, and a second portion consisting of the remaining posts, which was evaluated by a single annotator. The results of the evaluation campaign showed significant, consistent improvement of translation quality due to pre-editing, for all the domains and language pairs considered in the project. However, at the time Deliverable D9.2.2 was submitted, for the English-German language pair only results for the first data portion were available. The investigation of the second data portion was performed during Year 3 of the project. Additional work was carried out for German as a target language in two other directions: i) the analysis of the degraded sentences, and ii) the automatic evaluation of the impact of pre-editing rules on translation quality, in order to study the correlation between human and automatic evaluation results. Below, we describe the evaluation methodology, the experimental setup and the results obtained.

Evaluation Methodology

The human evaluation of the impact of ACCEPT pre-editing rules on SMT quality followed the methodology detailed in Deliverables D2.1 and D9.2.2. To summarise, each evaluation unit (forum post, in our case) is pre-edited[3] and machine translated, after which the translations corresponding to the original and pre-edited source versions are shown to evaluators in random order for comparative ranking. The ranking is based on a 5-point scale (*first clearly better, first slightly better, about the same, second slightly better, second clearly better*). This scale is then reduced to a 3-point scale for subsequent statistical analysis (*first better, about the same, second better*). The evaluators have access to the source sentence for reference.

Experimental Setup

Table 1 below summarises the experimental setup. (For the sake of clarity, we will use the same type of presentation for all experiments described in this deliverable.)

Experimental setup	
Language pair	English-German
Domain	IT technical forum ("Symantec")
Rule set	Preediting_SMT_Eval
Checking options[4]	1. Automatic {+spelling,+grammar,+style} 2. Manual {+spelling,+grammar,+style}
MT system	ACCEPT baseline[5]
Evaluation unit	forum post

Table 1: Evaluation of pre-editing impact for English-German – Experimental setup

[2] Deliverable D9.2.2 also includes evaluation results for the healthcare data domain. In the present deliverable, we only considered the technical forum domain, since this was the focus of recent developments in the project.
[3] The pre-editing process consists in the sequential application of the entire rule sets defined in Deliverable D2.2 for the Symantec domain. For French, there are three rules sets: Portal_Set_1, Portal_Set_2, Portal_Set_3, designed for automatic, manual and silent application, respectively (see Appendix 1 of Deliverable D2.2). For English, there is a single rule set, Preediting_SMT_Eval, applied first in automatic, then in manual mode (see Appendix 2 of Deliverable D2.2).
[4] Checking options control the activation and deactivation of rules of specific types (spelling, grammar, style). The plus sign denotes that rules of the specified type are activated, and the minus sign denotes that they are deactivated.
[5] By "ACCEPT baseline" MT system, we mean the ACCEPT baseline Symantec system described in Deliverable D4.1. This holds for all mentions in this deliverable.

One evaluator participated in this evaluation task. Like the other evaluators in the evaluation campaign, this evaluator was an advanced Master's student in translation, a native speaker of German with working knowledge of English, with no particular expertise in the domain. The evaluator received guidelines[6] and was paid for the task.

Evaluation Results

To report the performance of the ACCEPT pre-editing rules for the English-German language pair, we take into account the judgements elicited from the evaluator for the complete dataset (including both portions). The new results complete the Table 5 from Deliverable D9.2.2, which reported the impact of pre-editing on translation quality for the other language pairs, French-English and English-French. Table 2 below shows the complete results. The same information is presented graphically in Figure 1.

Impact	French-English	English-French	English-German
better	53.9%	49.8%	58.4%
same	30.0%	23.1%	9.4%
worse	16.1%	27.1%	32.2%
N	1756	1569	1598

Table 2: Pre-editing impact for all language pairs and all data[7]

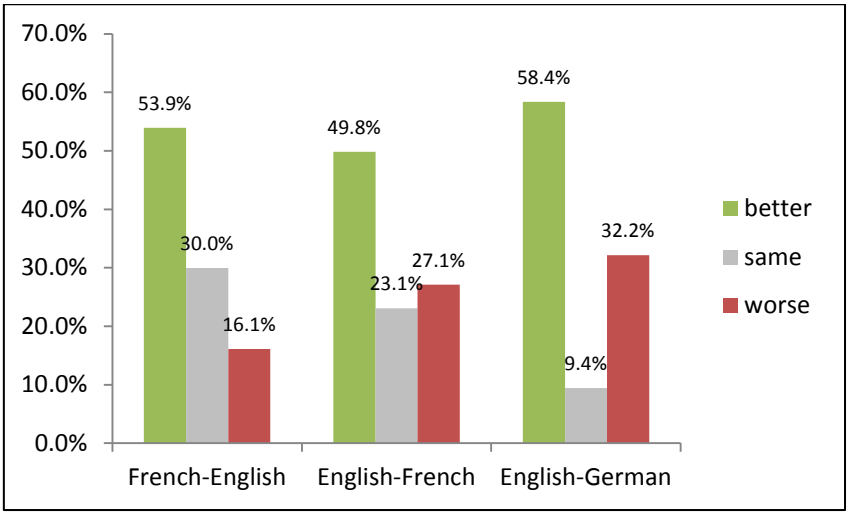

Figure 1: Pre-editing impact for all language pairs and all data

Statistical Significance

As with the other language pairs used in the project (see Deliverable D9.2.2), a McNemar test was conducted for the English-German language pair in order to compare the number of cases in which the translation became better vs. worse due to pre-editing. The difference is *extremely statistically significant*, $p < 0.0001$, in line with results for the other languages.

[6] The evaluation guidelines can be found at http://www.accept.unige.ch/Products/D9.2.4-Manual-Evaluation-Guidelines.pdf (Accessed: December 2014).

[7] N = number of posts in the dataset whose translation was affected by pre-editing.

Analysis of Degraded Sentences

Further evaluation work has been carried out for the English-German language pair in order to understand why pre-editing sometimes leads to worse translation. To this end, a human judge performed an analysis of degraded sentences, as described in the task guidelines.[8] The data analysed was a subset of the first portion of the dataset. From the total of 500 posts, 63 were selected for analysis. The selection criterion was that a negative impact of pre-editing was unanimously observed for these posts by the evaluators (i.e., these posts were judged as "worse" by all three evaluators).

The annotator was instructed to compare the two translation versions, to focus on the translation which was worse, and to identify the main cause why this translation was judged worse than the other. They were asked to come up with a classification of errors introduced by pre-editing rules, e.g. "wrong capitalization", "wrong word choice", "wrong number", "wrong tense" or "grammar error".

The classification provided by the annotator has been adapted slightly to group the classes clearly targeting the same phenomenon (e.g., "grammar error" and "syntactic error" were grouped into a "syntax" class; "wrong program name" and "wrong user name" were grouped into a "proper name" class; "wrong article" and "wrong word choice" were grouped into a "lexical choice" class). The final classification and the error distribution according to this classification are shown in Figure 2.

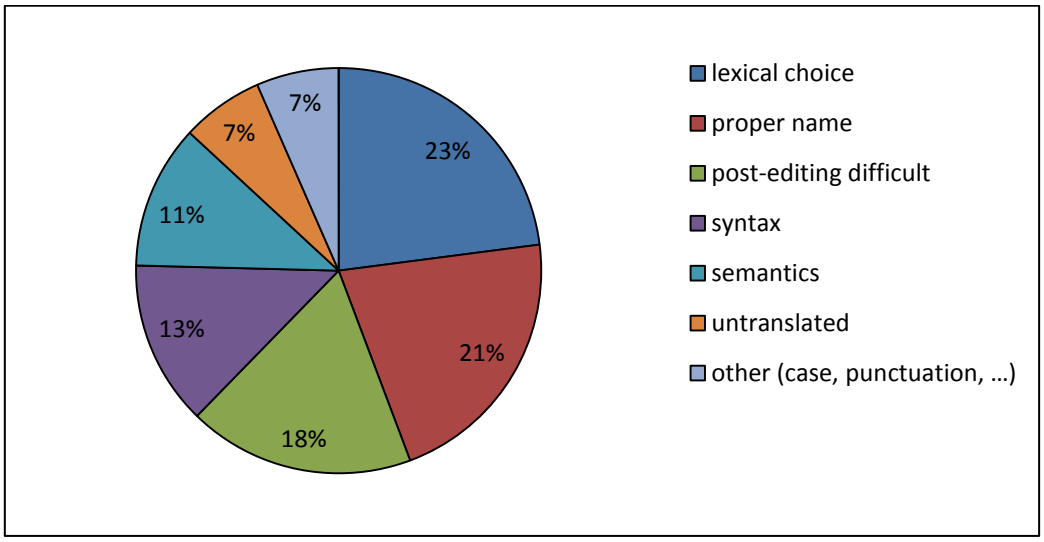

Figure 2: Error distribution for English-German

The results of the analysis showed that a high number of errors are due to automatic spelling correction of proper nouns. The identification of proper nouns is a well-known problem for user-generated content (Bontcheva et al., 2013). In subsequent evaluation experiments, spelling correction was performed manually, as it was integrated into the manual pre-editing rule set.

[8] The task guidelines can be found at http://www.accept.unige.ch/Products/D9.2.4-Error-Analysis-Guidelines.pdf (Accessed: December 2014).

2.2 Correlation between Automatic and Human Comparative Evaluation Results

The performance of ACCEPT pre-editing rules is assessed in WP9 in both an automatic and a manual way, the objective automatic scoring mechanism balancing the subjective human evaluations.

Previous results of the automatic evaluation using metrics like BLEU, GTM, METEOR and TER (D9.2.2) showed that the impact of pre-editing rules is not well reflected by metric scores (at the document level). To summarise, the evaluation was performed on a subset of 50 French forum posts randomly selected from the total 2000 posts used in the comparative evaluation campaign (see Section 2.1). Document-level scores were computed using reference translations into English produced from scratch by a semi-professional translator (an advanced Master's student in translation, a native speaker of the target language with no particular expertise in the domain).[9] The score difference due to pre-editing was not statistically significant, according to the paired t-test ($p > 0.05$). The largest impact was observed in the TER metric scores, $t(49) = -1.680$, $p < 0.1$. The correlation between human annotations and automatic scores results was positive but *weak* (Kendall's tau between 0.130 and 0.211), in line with reports from literature. The highest correlation result was achieved by the METEOR metric (Kendall's tau = 0.211, $p = 0.056$). However, none of the metrics correlated significantly with human annotations.

In Year 3 of the project, new experiments have been carried out to further investigate the issue of correlation between automatic and human evaluation results. The first experiment studied the correlation at the sentence level, as opposed to document level as in previous work. The second experiment focused on the English-German language pair and studied the correlation at the document level. The two experiments are described below.

Correlation at the Sentence Level (French-English)

As discussed above, the results of previous work (see D9.2.2 for details) showed weak correlation between human ratings and metric scores when evaluating the impact of pre-editing on translation quality. In that experiment, the correlation was computed at the document level, as the evaluation unit was the forum post (see also Section 2.1). Each post had a corresponding human rating, from the 3-point scale *better-same-worse*, and an automatic rating, from the same scale, which was obtained by taking into account the difference in raw metric scores (positive, null, negative). This transformation allowed the computation of the correlation coefficient between human and automatic ratings.

In the new experiment, our aim was to find out if a better correlation can be obtained when we consider a finer granularity for the data. We therefore elicited human judgements and computed metric scores at the sentence level. The research question addressed was the following:

(RQ1) When the source text is pre-edited, does the change in translation quality as rated by human evaluators correlate at the sentence level with the change observed in metric scores?

The experimental setup is summarised in Table 3.

[9] The translator received guidelines and was paid for the task. The translation guidelines are included in Appendix B.3 of deliverable D9.2.2.

Experimental setup	
Language pair	French-English
Domain	IT technical forum (“Symantec”)
Rule set	{Portal_Set_1, Portal_Set_2, Portal_Set_3}
Checking options	1. Automatic {+spelling,+grammar,+style} 2. Manual {+spelling,+grammar,+style} 3. Silent {+spelling,+grammar,+style}
MT system	ACCEPT baseline
Evaluation unit	sentence

Table 3: Correlation at the sentence level – Experimental setup

The data from this experiment is the same as in the previous experiment, namely the 50 forum posts pre-edited using the ACCEPT rules (including automatic spelling correction). The posts were manually split into sentences. From the 224 resulting sentences, those having identical translations for the raw and pre-edited versions were removed. This left us with 107 sentences that were used in the experiment. (Note that manual splitting into sentences was necessary, because sentence boundaries are often missing in user-generated content, making automatic segmentation problematic.)

For the 107 sentences in the dataset, we collected human judgements using the same evaluation method as in the large comparative evaluation campaign (see Section 2.1). The evaluators were two ACCEPT collaborators, native/near-native speakers of the target language. The inter-annotator agreement is *moderate* (Cohen’s k = 54.9%). For subsequent analysis, the disagreement cases have been settled as follows:

- Disagreements between *better* and *worse* categories (6/107 cases) were settled by letting the annotators discuss conflicting ratings until they reached agreement;
- Disagreements between *same* and another category were not considered as real disagreements, but as reluctance of one of the judges to take a stance. We therefore considered the other’s judge opinion as the significant one and used their category in our analysis.

After cases involving disagreement were settled, the comparative evaluation results showed the following distribution (Table 4):

Impact	French-English
better	57.0%
same	20.6%
worse	22.4%
N	107

Table 4: Pre-editing impact at the sentence level

Automatic metric scores were computed for the 107 sentences in the dataset and, based on these scores, positive/null/negative labels were assigned to each sentence, according to the observed impact on metric score due to pre-editing (similarly to human evaluation). The reference translations required for score computation were extracted from the original set of 50 reference post translations by manual alignment of sentences.

We computed the correlation between human and score labels and found statistically significant *weak* correlation between the BLEU metric and human annotations (Kendall's tau = 0.263, $p < 0.01$), as well as between the TER metric and human annotations (Kendall's tau = 0.210, $p < 0.05$). The GTM and METEOR metric did not correlate significantly with human ratings. The results are shown in Table 5.

Metric	Kendall's tau
BLEU	0.263
GTM	0.162
METEOR	0.155
TER	0.210

Table 5: Correlation at the sentence level - Results

Hence, we can conclude that human ratings of pre-editing impact on translation quality correlate weakly and significantly with two of the metrics investigated (BLEU and TER), but there is no significant correlation with the other metrics (METEOR and GTM). The weak correlation results are in line with the earlier experiment outlined at the beginning of this section. From the point of view of statistical significance, the new results however contrast with those of the earlier experiment, performed at the document level, which showed no significant correlation between human ratings and metric scores. This contrast is most plausibly explained by the fact that the present study involved a larger number of observations, thanks to the increased granularity of the data.

Correlation at the Document Level (English-German)

As explained in Section 2.1, new human evaluation results have been made available for the English-German language pair. To further investigate the issue of correlation between human ratings and automatic metric scores, a new experiment was carried out, using a methodology similar to that in the first (document level) experiment for French-English.

In the new experiment, our aim was to find out whether automatic metric scores and human judgements correlate, and also to identify the best-performing metric in this context (English user-generated content pre-edited and translated with an SMT system to German). The research question addressed was the following:

(RQ2) When the source text is pre-edited, does the change in translation quality as rated by human evaluators correlate at the document level with the change observed in metric scores?

The experimental setup is summarised in Table 6.

Experimental setup	
Language pair	English-German
Domain	IT technical forum ("Symantec")
Rule set	Preediting_SMT_Eval
Checking options	1. Automatic {+spelling,+grammar,+style} 2. Manual {+spelling,+grammar,+style}
MT system	ACCEPT baseline
Evaluation unit	forum post

Table 6: Correlation at the document level – Experimental setup

The data in this experiment consisted of 50 English forum posts from the larger dataset of 2000 posts used in the Year 2 evaluation campaign (see Section 2.1). These posts had been previously pre-edited using the ACCEPT rules, which included automatic spelling correction. The 50 posts were randomly selected from the posts with many rule activation flags. More precisely, we retained only posts with a higher than average number of automatic corrections; since the average number of corrections for the whole collection of 2000 posts is 2.8, we only included posts with at least 3 corrections. As in the previous experiment for French-English, the posts were also selected such that they have a medium size (between 186 and 500 characters). The reason behind this criterion was that, according to classification work reported in Deliverable D3.1, posts of moderate length are overwhelmingly useful.

The 50 posts selected were translated into English from scratch by a semi-professional translator (advanced Master's student in translation, native speaker of the target language, with no particular expertise in the domain).[10] The difference in metric scores obtained due to pre-editing was statistically significant in the case of the METEOR metric ($t(49) = 2.158$, $p = 0.036$), but non-significant for the BLEU and TER metrics.

The Kendall's tau correlation coefficient was computed between 1) the human rating falling in the scale better-same-worse, and 2) the automatic metric document-level scores for the posts, transformed into a positive-null-negative scale, as in previous experiments. The correlation results are shown in Table 7.

Metric	Kendall's tau
BLEU	0.309
METEOR	0.375
TER	0.461

Table 7: Correlation at the document level - Results

The results of the experiment show a positive statistically significant correlation between metric scores and human annotations. The correlation remains in the *weak* range for BLEU and METEOR, but is *moderate* for the TER metric. The results are reliable ($p < 0.05$ in the case of BLEU; $p < 0.01$ in the case of TER and METEOR). The relatively higher correlation results obtained with respect to previous experiments may be explained by the use of a different sampling method, which favoured posts with many corrections.

The results reported in this section suggest that there is no single metric suitable for all tasks, but rather that the performance of metrics varies according to the experimental context.

2.3 Impact of ACCEPT Pre-editing Rules on a Different Data Domain

The impact of ACCEPT pre-editing rules has until now been evaluated on in-domain data, i.e., on a withheld portion of the data from the Norton Community forum, similar to the data that served for tuning the ACCEPT baseline SMT system and for developing the pre-editing rules. To extend the evaluation of rules, we performed an experiment with data from a different domain, using a general-purpose MT system, Google Translate.[11] This experiment is presented in detail in Gerlach (2015). The experimental setup is summarised in Table 8.

[10] The translator received guidelines and was paid for the task. The translation guidelines can be found at http://www.accept.unige.ch/Products/D9.2.4-Translation-Guidelines.pdf (Accessed: December 2014).
[11] https://translate.google.com/ (Accessed: July 2014).

Experimental setup	
Language pair	French-English
Domain	DIY technical forum ("Plumbing")
Rule set	{Portal_Set_1, Portal_Set_2, Portal_Set_3}
Checking options	1. Automatic {-spelling,+grammar,+style} 2. Manual {+spelling,+grammar,+style} 3. Silent {+spelling,+grammar,+style}
MT system	Google Translate
Evaluation unit	sentence

Table 8: Impact of pre-editing rules on a different domain – Experimental setup

The dataset consisted of 1000 sentences extracted from posts on plumbing taken from a French-language DIY forum.[12] The sentences were pre-edited using the ACCEPT automatic and manual rules defined for the Symantec technical forum domain, with manual application of spelling correction rules. They were translated with the Google Translate MT system, and finally evaluated by three human judges. For subsequent analysis, only the source sentences changed by pre-editing were retained, i.e., 751 sentences (75.1%). In a further 132 cases, pre-editing had no impact on translation, which left us with 619 sentences to evaluate.

The human comparative evaluation methodology was the same as the one used in the previous evaluation campaign and in the experiments reported in this deliverable (see Section 2.1), except that the judges were non-experts users instead of translation professionals. They were recruited and performed their work on the Amazon Mechanical Turk platform. They are Canadian, self-declared bilingual speakers of French and English, and were paid for the task.

The inter-annotator agreement was *fair* (Fleiss' k = 0.386).[13] The subsequent analysis was based on majority judgements, i.e., on the cases in which at least two of the judges agreed when the transformed scale better-same-worse was taken into account. The cases in which no majority label could be assigned have been left aside (31 sentences).

Table 9 reports the evaluation results on the remaining 588 sentences. As can be seen, these results are comparable with the results achieved for in-domain data for the same language pair (see Deliverable D9.2.2, Table 5). These findings show that the ACCEPT pre-editing rules are portable to a radically different data domain.

Impact	French-English
better	69.4%
same	7.5%
worse	23.1%
N	588

Table 9: Impact of pre-editing rules on a different domain - Results

[12] http://www.bricoleurdudimanche.com/fiches-bricolage/plomberie-et-sanitaires-75.html (Accessed: July 2014).

[13] The Fleiss' kappa measure (Fleiss, 1981) is an inter-annotator agreement measure which is appropriate for experiments where there are more than two annotators (coders), and these are randomly selected from a larger population of coders (Hallgren, 2012).

For consistency with previous work, we conducted a McNemar test to compare the number of cases in which the translation became better vs. worse due to pre-editing in the new domain. The difference is *extremely statistically significant*, $p < 0.0001$, in line with previous results.

Further similar work was carried out by Gerlach (2015) on data which is only slightly out of domain, namely, sentences from the French IT technical forum CNET.[14] The results are in the same range as the results on in-domain and out-of-domain data reported in Gerlach (2015) and comparatively summarised here in Table 10. Note that, in the three cases, the MT system used was Google Translate. The number of rule activation flags is relatively similar for in-domain and slightly out-of-domain data, and relatively lower for out-of-domain data, as can be seen in the second column of Table 10.

Data type	Rule flags/100 words	Source changed	Target changed	N	Impact of pre-editing on translation quality		
					better	same	worse
Out of domain (DIY)	11.8	751	619	588	69.4%	7.5%	23.1%
Slightly out of domain (CNET)	13.5	697	486	466	83.7%	3.4%	12.9%
In domain (Symantec)	13.3	665	534	508	74.0%	3.3%	22.6%

Table 10: Impact of pre-editing rules on a different domain – Results of related work

3. Performance of ACCEPT Pre-editing Rules across MT Paradigms

In this section, we describe the experiments carried out to compare the performance of ACCEPT pre-editing rules achieved in an SMT scenario versus in an RMBT scenario.

The first experiment was devoted to comparing the impact of ACCEPT pre-editing rules across MT paradigms by investigating the performance at the level of individual rules (that is, rules like *subject-verb agreement* or *wrong capitalisation* were evaluated separately). The second experiment investigated the performance of rules in a global manner. All triggered rules were applied on the source text, then rules were evaluated in combination (i.e., their combined impact was evaluated). In the third experiment, we performed the evaluation of the combined impact of rules at the document level, as opposed to the sentence-level as in the two previous experiments.

3.1 Performance of Individual Rules

This experiment was designed to test the portability of ACCEPT pre-editing rules to an RBMT setting and to compare the performance of individual rules across paradigms (SMT vs. RBMT). The experiment is detailed in Gerlach (2015).

The research question addressed was the following:

(RQ3) How does the individual performance of pre-editing rules compare between paradigms (SMT vs RBMT)?

The experimental setup is summarised in Table 11.

[14] http://forums.cnetfrance.fr/forum/51-securite/ (Accessed: July 2014).

Experimental setup	
Language pair	French-English
Domain	IT technical forum ("Symantec")
Rule set	{Portal_Set_1, Portal_Set_2, Portal_Set_3}
Checking options	1. Automatic {-spelling,+grammar,+style} 2. Manual {+spelling,+grammar,+style} 3. Silent {+spelling,+grammar,+style}
MT system	ACCEPT baseline, $Lucy_{specialised}$
Evaluation unit	sentence

Table 11: Impact of pre-editing across paradigms – Experimental setup

In order to make the comparison possible, the Lucy RBMT system was specialised to the ACCEPT project data domain. Its source, target and transfer lexica were extended to include Symantec terminology. This process was performed in three steps. First, a glossary of around 5000 Symantec product names was imported into the Lucy LT Lexshop 2.7 dictionary editor to create user dictionaries. Second, the Symantec translation memory (see Deliverable D4.1) was mined with the Lucy LT Desktop Power 2.1 translation engine to extract words unknown to the Lucy system; the 500 commonest were then added to the user dictionaries. Third, a similar process was followed to extract unknown words and their translations from the tuning data used for deploying the ACCEPT baseline system (see Deliverable D4.1).

The data in the experiment consisted of 1808 sentence extracted from French forum posts made available by the project partner, Symantec. They were sampled in such a way as to prefer inclusion of representative n-grams in the Symantec data. (The same dataset had been used for a previous study on the impact of pre-editing on the ACCEPT baseline, described in Gerlach et al. 2013).

For the purpose of this experiment, the dataset was translated with the Lucy system specialised for the project domain, and the translation pairs were evaluated by three human judges using the standard evaluation methodology (outlined in Section 2.1). The judges were Amazon Mechanical Turk workers, Canadian, self-declared bilingual speakers of French and English. They were paid for the task.

Table 12 presents the results in terms of impact of pre-editing in the RBMT and SMT settings. As can be seen, the results are very similar across paradigms, which means that the ACCEPT pre-editing rules, originally tailored to an SMT setting, can be successfully ported to an RBMT setting.

Impact	SMT	RBMT
better	72.7%	71.2%
same	8.6%	7.0%
worse	18.7%	21.8%
N	1362	1376

Table 12: Impact of pre-editing across paradigms – Results[15]

[15] N = number of sentences on which pre-editing had an impact and to which a majority label could be assigned.

The Spearman's rank correlation coefficient was computed to find out if the best performing rules are the same in the two settings. The results showed a positive *weak* correlation between the two settings (Spearman's rho = 0.397, $p < 001$). The correlation varies according to the rule type (i.e., whether it is has been designed for automatic, manual, or silent application; see Deliverable D2.2). The correlation results by rule type are shown in Table 13. The ANOVA analysis of variance showed, however, that the effect of rule type was not significant, $F(2, 65) = 0.543$, $p = 0.583$. This means that all rule types are ultimately equally performant in both paradigms.

Rule Set	Spearman's rho
automatic	0.517 (p < 0.001)
manual	0.359 (N.S.)
silent	0.263 (N.S.)

Table 13: Correlation of rule performance across paradigms according to rule type

Further investigation revealed that some of the rules are more effective in the RBMT setting than in the SMT setting (e.g., *erreur_de_majuscule, homophones_verbe_nom, accord_sujet_verbe*), and vice versa (e.g., *mettez_impératif, évitez_est_ce_que, evitez_tu*). However, overall the difference in rule performance across paradigms is not statistically significant, which confirms the portability of rules.

When comparing the number of cases in which translation became better vs. worse due to pre-editing, in both settings the difference is *extremely statistically significant* according to the McNemar test ($p < 0.0001$). This result is in line with the results of similar work reported in this deliverable.

3.2 Combined Impact of Rules at the Sentence Level

The aim of the second experiment was to test the portability of ACCEPT pre-editing rules to an RBMT setting, by investigating the combined impact of rules, rather than the impact of individual rules as in the previous experiment (Section 3.1).

The research question addressed was the following:

(RQ4) What is the combined impact of pre-editing rules in an RBMT setting?

The experimental setup is summarised in Table 14.

Experimental setup	
Language pair	French-English
Domain	IT technical forum ("Symantec")
Rule set	{Portal_Set_1, Portal_Set_2, Portal_Set_3}
Checking options	1. Automatic {-spelling,+grammar,+style} 2. Manual {+spelling,+grammar,+style} 3. Silent {+spelling,+grammar,+style}
MT system	$Lucy_{specialised}$
Evaluation unit	sentence

Table 14: Combined impact of rules (sentence level) – Experimental setup

The data in this experiment consisted of 100 French forum posts from the larger dataset of 2000 posts used in the Year 2 evaluation campaign (see Section 2.1). The posts were randomly selected from the useful posts, according to the usefulness criteria defined in Deliverable D3.1 (see also Section 2.2).

The posts were pre-edited with the three pre-editing sets defined for French in WP2 (see Deliverable D2.2). The first set of automatic rules was applied with spelling checking turned off. The second set of manual rules – including spelling correction rules – was applied by one of the authors in interactive mode, using the Acrolinx Word plug-in. The third set of silent rules specifically tailored for machine translation was applied last, in an automatic way.

After pre-editing, the posts were automatically translated with the Lucy RBMT system specialised for the ACCEPT data domain (described in Section 3.1). The posts were then automatically split into sentences. Of the 503 resulting sentences, 292 had identical translations when pre-edited (in 190 of these, the source was not changed by pre-editing).

The remaining 211 sentences underwent comparative evaluation by three Amazon Mechanical Turk workers, Canadian, self-declared bilingual in French and English, paid for the task. The evaluation methodology was the same as in the case of the first experiment (Section 3.1). The inter-annotator agreement was *fair* (Fleiss' k = 0.309). Table 15 reports the evaluation results for the data where a majority label could be assigned, i.e., 205 sentences.

Impact	SMT
better	71.7%
same	9.8%
worse	19.5%
N	205

Table 15: Combined impact of rules (sentence level) – Results[16]

As can be seen, the combined impact of pre-editing rules in the RBMT setting is very similar to the impact observed at the level of individual rules in both the RBMT and SMT setting (see Table 12), and in line with other results reported in this deliverable in relation to rule performance.

According to the McNemar test, the difference in the number of cases in which translation became better vs. worse due to pre-editing is *extremely statistically significant* ($p < 0.0001$).

A similar experiment has been carried out by Gerlach (2015) on a sample of 1000 sentences from Norton Community forum posts, which were processed and evaluated using the same procedure as above. The impact of pre-editing on the output of four MT systems investigated was found similar to the impact observed in the present experiment, as can be seen from Table 16.

[16] N = number of sentences on which pre-editing had an impact and to which a majority label could be assigned.

System	Source changed	Target changed	N	Impact of pre-editing on translation quality		
				better	same	worse
ACCEPT baseline	665	533	520	82.9%	6.0%	11.2%
$Lucy_{specialised}$	665	526	508	84.8%	4.3%	10.8%
$Systran_{specialised}$[17]	665	483	464	86.9%	0.4%	12.7%
Google Translate	665	534	508	74.0%	3.3%	22.6%

Table 16: Combined impact of rules (sentence level) – Results of related work

3.3 Combined Impact of Rules at the Document Level

Similarly to the work reported above, the next experiment tests the portability of ACCEPT pre-editing rules to an RBMT setting, but this time by taking into account the combined impact of rules at the document (forum post) level.

The research question addressed was the following:

(RQ5) How does the combined impact of pre-editing rules compare between paradigms (SMT vs RBMT) at the document level?

The experimental setup is summarised in Table 17.

Experimental setup	
Language pair	French-English
Domain	IT technical forum ("Symantec")
Rule set	{Portal_Set_1, Portal_Set_2, Portal_Set_3}
Checking options	1. Automatic {-spelling,+grammar,+style} 2. Manual {+spelling,+grammar,+style} 3. Silent {+spelling,+grammar,+style}
MT system	ACCEPT baseline, $Lucy_{specialised}$, $Systran_{specialised}$
Evaluation unit	forum post

Table 17: Combined impact of rules (document level) – Experimental setup

The data in this experiment consists of the 100 French forum posts from the previous evaluation experiment performed at the sentence level (Section 3.2). The posts and their pre-edited version were translated using three MT systems, one instantiating the SMT paradigm (i.e. the ACCEPT Symantec baseline system)[18] and two instantiating the RBMT paradigm (i.e. the Lucy and Systran systems specialised for the ACCEPT project data domain). The sentences for which pre-editing had an impact were retained for human evaluation. The final test sets consist of 94 posts in the case of the ACCEPT baseline system, 96 posts in the case of Lucy, and 91 in the case of Systran.

For each system, the pairs of translation versions were comparatively evaluated by Amazon Mechanical Turk workers using a 5-point Likert scale, which was subsequently converted into a 3-point scale *better-same-worse* (according to the standard evaluation methodology presented in Section 2.1). As before, the judges were randomly chosen from Canadian workers, who were self-declared bilinguals. Their task was to compare the translations of the pre-edited and of the original versions, shown in random order, while having access to the pre-edited version of the source. The Amazon Mechanical Turk workers (henceforth, for simplicity, users) were paid for the task.

[17] The specialisation of the Systran system was performed similarly to that of the Lucy system (Section 3.1). The version 7.3.5.13 of the system was used, and more precisely, its RBMT functionality (not the hybrid one).
[18] This system is described in Deliverable D4.1.

The same evaluation method was used to collect ratings from one expert (one of the authors). These ratings are not taken into account in the subsequent analysis, but are reported here solely for the purpose of comparison with user ratings, in order to assess the reliability of the latter (cf. Task 9.3 of the DOW).

The inter-annotator agreement statistics are shown in Table 18. The results show *moderate*, close to *strong* agreement between users, indicating that their ratings can be reliably used for subsequent statistical analysis. In order to compute the agreement between user ratings and expert ratings, we took into account the majority label assigned by users to each post. Disagreement cases (i.e. label combinations *better-same-worse*) were settled by using the *same* label. The agreement statistics show that there is a very high observed agreement between expert and non-expert ratings, while chance-corrected agreement is only *fair* to *moderate* (the low Cohen's kappa is explained by the prevalence in the distribution of the data, cf. Byrt et al., 1993). The Spearman's correlation coefficient is very high and indicates significant, *moderate* to *strong* correlation between expert and non-expert rating in the three scenarios ($p < 0.01$). These findings are in line with results from our previous work, which also showed significant, strong correlation between user and translator judgements (cf. Section 3 of Deliverable D9.2.2).

MT system	Fleiss' k (users)	Expert-users agreement		
		Observed agreement	Cohen's k	Spearman's rho
SMT: ACCEPT baseline	0.547	0.723	0.384	0.539
RBMT: Lucy	0.596	0.789	0.390	0.569
RBMT: Systran	0.433	0.747	0.412	0.568

Table 18: Combined impact of rules (document level) – Inter-annotator agreement statistics

On the basis of the users' ratings for the three test sets, we computed the impact of pre-editing rules across paradigms (reported in Table 19 below).

Impact	SMT: ACCEPT baseline	RBMT: Lucy	RBMT: Systran
better	83.9%	88.4%	85.2%
same	4.3%	2.1%	3.4%
worse	11.8%	9.5%	11.4%
N	93	95	88

Table 19: Combined impact of rules (document level) – Results[19]

As can be seen, the ACCEPT pre-editing rules achieve a very similar performance across paradigms. In an overwhelming number of cases, pre-editing has a positive impact on translation quality. The results are better than the results of similar evaluations performed in the Year 2 campaign (see Deliverable D9.2.2, Table 5). This finding can be explained by the application of spelling correction rules in interactive, rather than automatic, mode (see note in Section 2.1). Another explanation is the use of a different data sampling strategy, targeting medium-size 'useful' post as opposed to random posts.

[19] N = number of posts on which pre-editing had an impact and to which a majority label could be assigned.

In terms of statistical significance, for each system the difference in the number of cases in which translation became better vs. worse due to pre-editing is *extremely statistically significant* according to the McNemar test ($p < 0.0001$).

4. Performance of ACCEPT Post-editing Rules

In this section, we describe the work carried out to evaluate the ACCEPT post-editing rules released in the second half of Year 3 of the project (see Deliverable D2.4).

The individual performance of post-editing rules has previously been measured and reported in Deliverable D2.4. Below, we report on an experiment designed to evaluate rules in a global way, by looking at the combined impact of post-editing rules on translation quality. This experiment is detailed in Porro et al. (2014).

The aim of the experiment was to quantify the extent to which the automatic correction of frequent errors in MT output (as provided by post-editing rules) contributes to improving the output, before the latter is submitted to human post-editing. We therefore elicited comparative judgements from language professionals relating to the impact of automatic post-editing on the *readability* of the MT output. It might be argued, however, that the MT output may be improved in a way that is irrelevant from a human post-editing perspective. We therefore complemented the readability evaluation with a further analysis, in which we took into account the *usefulness* of the automatic changes introduced by rules. We let the language professionals further post-edit the text manually, then looked at whether the automatic changes were useful, i.e. were kept in the final version produced by post-editors.

The experiment addressed two main research questions:

(RQ6) When the target text is automatically post-edited, what is the impact on text readability of the changes introduced?

(RQ7) When the target text is automatically post-edited, are the changes introduced useful, i.e. are they actually kept in subsequent manual post-editing?

The experimental setup is summarised in Table 20.

Experimental setup	
Language pair	English-French
Domain	IT technical forum (“Symantec”)
Rule set	Postediting-EN-FR
Checking options	Automatic {-spelling,+grammar,-style}
MT system	ACCEPT baseline
Evaluation unit	sentence

Table 20: Impact of post-editing rules – Experimental setup

The experimental data consists of about 5000 pre-edited sentences corresponding to the 2000 English forum posts used in the comparative evaluation campaign (see Section 2.1). The translations obtained using the ACCEPT baseline MT system were automatically post-edited using the French post-editing rules defined in WP2 (see Deliverable D2.4).

In order to better target the evaluation effort, we discarded the sentences which were very long (more than 40 words), as well as the sentences with a very low number of changes introduced by post-editing rules (low Levenshtein distance between the corrected and non-corrected MT output versions). The 200 sentences with the highest number of changes were considered for investigation. After removing a duplicate sentence, we ended up with a final test set of 199 sentences.

The test set was comparatively evaluated in terms of impact on readability by three advanced Masters' students in translation, native French speakers with English as their main working language. The participants did not have particular expertise in the domain. They received task guidelines and were paid for the task.

The evaluators comparatively rated each translation pair on a 3-point scale *first better – same – second better* (for consistency with our standard evaluation methodology). The order of original and post-edited translation versions was randomly swapped. In contrast to previous work, the evaluators in this experiment did not have access to the source sentences. The rationale behind this methodological choice was that the evaluation was focused only on readability, with no consideration of adequacy.

In the same task, in addition to the global rating for a pair of sentences, the evaluators provided ratings for the individual changes introduced by rules, using the same scale as before. A total of 391 individual changes were thus evaluated in terms of impact on readability by the three evaluators.

The results on the impact of post-editing rules on text readability are shown in Table 21.

Readability	Sentence level	Individual change level
better	82.0%	82.8%
same	17.0%	14.8%
worse	1.0%	2.3%
N	194	384

Table 21: Impact of post-editing rules – Readability results[20]

The results show that both at the sentence and individual change level, the impact of automatic post-editing rules on text readability is overwhelmingly positive.

To answer the second research question, a post-editing experiment was carried out with the same participants who took part in the readability evaluation experiment. The post-editing task was performed using the ACCEPT post-editing environment (described in Deliverable D5.6). The post-editors had access to the pre-edited source and were asked to render the MT output grammatical and conformant to the meaning of the source sentence, while using as much of the MT output as possible (style was not given priority). They were provided with task guidelines and a glossary for the domain. The participants were paid for the task.

The manually post-edited output was compared against the automatically post-edited output to check whether individual automatic changes were kept during manual post-editing. The 391 individual changes introduced by rules were labelled as *found* or *missing*, accordingly. The usefulness

[20] N = number of items (sentences/individual changes) to which a majority label could be assigned.

is computed as the percentage of individual changes found in the manually post-edited text. The results are shown in Table 22.

Usefulness	Individual Change Level
found	71.9%
missing	28.1%
N	391

Table 22: Impact of post-editing rules – Usefulness results[21]

The correlation between readability and usefulness in the evaluation results is positive but *weak* (Kendall's tau = 0.307, $p < 0.01$). This means that the text readability and the usefulness from a post-editing perspective are distinct dimensions of the target text, which do not strongly overlap. As a matter of fact, local improvements at the level of grammar (e.g., fixing a verb tense) are irrelevant if the larger context is badly translated and the MT output has to be rewritten by post-editors. Consequently, both readability and usefulness dimensions have to be taken into account when evaluating the impact of post-editing rules.

To quantify the extent to which the automatically corrected MT output becomes more similar to the human post-editing output (gold standard) with respect to the original MT output, we computed the difference between each MT output version and the gold standard. In terms of TER, the difference is smaller (0.27) for the corrected output, and larger (0.42) for the original output. This confirms that the post-editing rules indeed help make the MT output more similar to the human output.[22]

When the gold standard consists of reference translations produced from scratch rather than by post-editing, the automatically corrected MT output is also more similar to the gold standard (TER: 0.59) than the original MT output is (TER: 0.66).

Taken together, these findings on readability, usefulness and similarity to reference translation show that the ACCEPT automatic post-editing rules are beneficial for subsequent human post-editing, as they perform useful modifications which reduce the number of changes the post-editors have to make in order to reach the final output.

The findings of the present experiment are reliable, the chi-square goodness of fit test proving that the difference in proportions observed in both readability and usefulness results is statistically significant.

5. User Interaction with Pre-editing Rules

In Task 9.6, we have developed methods to adapt the application of pre-editing rules based on feedback from users of these rules. More specifically, the aim was to develop tools that help continuously monitor and analyse how ACCEPT users work with pre-editing rules in practice, and to identify potential problems with specific rules.

In contrast to dedicated user studies with volunteers conducted in WP7 and WP8, these methods provided insights into the behaviour of "average" forum users while they were pre-editing their own content. This section describes the collected results for users of the English and French Norton

[21] N = number of items (individual changes) to which a majority label could be assigned.
[22] The reviewer raised the question of human output quality. Our manual investigation of the three versions of post-edited output confirmed that the output is of reliable quality.

Community forum. This work also complements the analysis described in Section 3.5 of Deliverable D6.5. (Please refer to Deliverable D9.3 for details of the technical implementation of the data collection and aggregation mechanisms.)

5.1 Collected Datasets

To aggregate the user feedback, we collected all available usage data that had been logged by the ACCEPT pre-editing plug-in via the ACCEPT API. This data is collected in an anonymised form which makes it impossible to connect text items to any specific individuals using the Norton forum. More precisely, we accessed two datasets:

1. All available usage data from the English Norton Community Forum. This data ranges from June 2013, when the plug-in became able to record usage data, to September 2014, when the plug-in integration was disabled. The data applies to the English rule set "Preediting_Forum" that we created for interactive pre-editing in the English forum. The rule set contains the English rules for the forum use case presented in Deliverable D2.2.
2. All available usage data from the French Norton Community Forum, which ranges from November 2013, when the plug-in was integrated into the French forum, to September 2014, when the plug-in integration was disabled. The data applies to the French rule set "Preediting_Forum" that we created for manual pre-editing in the French forum. The rule set contains the French rules from Set1 and Set2 for the forum use case as described in Deliverable D2.2.

Note that the plug-in integration was disabled in September 2014 because the Norton Community forum at Symantec switched to different forum software. No resources were available to re-integrate and test the plug-in in the new forum software for the final months of the project. As the pre-editing plug-in was not integrated into the workflow used by Translators without Borders for the NGO use case, we could not collect usage data for that use case either. Nevertheless, we were able to collect a total of about 27700 individual data points for English and 15400 individual data points for French. We consider that these datasets are representative of user interaction with the pre-editing rules.

5.2 Collected Usage Data

Various forms of user interaction in the ACCEPT pre-editing plug-in are interpreted as implicit or explicit feedback on the flags and suggestions provided by Acrolinx. Table 23 summarises the recorded actions, their effect in the user interface, and their interpretation as user feedback.

The user actions on each flag were grouped by the rule that produced the flag. We also separately considered the two spelling flag types *general spelling flags*, which have suggestions of similarly written words from the spelling lexicon, and *unknown words*, which are spelling flags without such suggestions. We dropped all rules from the evaluation for which the dataset contained less than 15 occurrences of the relevant flags.

The actions where then aggregated into four main metrics:

- Flag attention: This metric describes the percentage of flags on which the user performed any observable interaction (actions 1 to 5 in Table 23);
- Flag precision: This metric describes the percentage of flags that were "accepted" by the user ("accept actions" are actions 1 and 5 in Table 23), out of those flags that the user eventually acted on;
- Suggestion proposal: This metric describes the percentage of flags for which the user was shown at least one replacement suggestion, out of those flags that the user accepted;

- Suggestion precision: This metric describes the percentage of flags for which the user selected a suggestion (action 1 in Table 23), out of those flags that where accepted by the user and that show at least one replacement suggestion.

	User action	Effect	Interpreted as feedback
1	Select a suggestion	The flagged text is replaced with the suggestion	The flag and the suggestion are accepted
2	"Ignore rule"	All the flags for this rule are removed now and in the future	The entire rule is rejected
3	"Learn Word"	Spelling flags on this word are removed now and in the future	Spelling flags on this word are rejected
4	Hover over flag for some time	A tooltip is shown	The flag is not sufficiently self-descriptive, and the user requires some help
5	Do nothing, change flagged text manually[1]	The flagged text is changed	The user accepts the flag, but none of the suggestions are considered suitable, or there are none
6	Do nothing, don't change flagged text[2]	The flagged text is unchanged	The user passively ignores this occurrence of the flag

Table 23: List of recorded user actions

5.3 Results for English Pre-editing rules

Table 24 shows the processed user feedback for the English pre-editing rules as obtained from dataset 1.

First of all, we noticed that the amount of flags that the forum users acted on ("flag attention") varies between 2% and 37%. The relatively low value may have been caused by users who were trying out the pre-editing functionality in the Norton forum, and acted only on a few or none of the flags before closing the pre-editing window.

The precision of the flags, i.e., the percentage of flags that led to a change of the text, was overall quite high. 11 out of the 21 considered rules had a precision of more than 80%, and another 4 rules were above 65%, which shows that the pre-editing rules in general proved to be quite useful.

Unsurprisingly, almost all of the rules with a low precision also had a very low attention. It is likely that the users "learned" that the rule was not useful, and just ignored it passively. This is most true for the "sentence too long" rule, which marks sentences it considers too long, but does not provide any suggestion. "Sentence too long" flags appeared often, but were often passively ignored; if the user acted on them, it was usually via the action "ignore rule". Hence the rule should probably be removed from the pre-editing rule set.

[1]Note: Whether the user changed the flagged text is detected heuristically by examining how the text looked like before and after the session.
[2]See note above.

Rule	Type	Flag count	Flag attention	Flag precision	Suggestion proposal	Suggestion precision
a/an distinction	grammar	77	16%	83%	90%	100%
duplicate punctuation mark	grammar	329	4%	36%	60%	100%
incorrect extra comma	grammar	219	3%	71%	0%	N/A
its/it is confusion	grammar	61	16%	80%	100%	100%
missing space	grammar	220	7%	73%	100%	100%
noun/adjective confusion	grammar	88	19%	100%	94%	94%
noun/adjective/verb confusion	grammar	76	14%	91%	100%	100%
number agreement	grammar	70	23%	88%	93%	100%
subject/verb agreement	grammar	206	17%	59%	85%	100%
use comma after introductory phrase	grammar	146	16%	78%	100%	100%
use comma after subordinate phrase	grammar	188	7%	43%	100%	100%
use end of sentence punctuation	grammar	147	18%	69%	100%	100%
write words together	grammar	27	37%	90%	100%	100%
wrong verb form	grammar	188	3%	67%	100%	100%
wrong word	grammar	32	19%	100%	100%	100%
sentence too long	style	1993	2%	3%	0%	N/A
avoid colloquialism and metaphorical language	style	180	10%	17%	100%	100%
general spelling flag	spelling	3434	24%	90%	98%	98%
capitalize at beginning of sentence	spelling	288	8%	100%	75%	75%
spelling error	spelling	164	18%	83%	92%	100%
unknown word	spelling	1356	2%	24%	0%	N/A

Table 24: Evaluation results for English pre-editing rules

Another rule with a low precision is “avoid colloquialism and metaphorical language”. While this rule may help the MT system in that it avoids phrases that are not in the phrase table and thus difficult to translate, it is not surprising that users are unfavourably disposed towards the idea that they should change typical aspects of forum language. Therefore, this rule should be revised.

The *general spelling flags* led to a user action in 24% of the cases, which is relatively good. Moreover, the action was to accept the flag in 90% of the cases, which shows that spell checking is a valuable feature in support forums.

If Acrolinx spell checker does not find similarly written words in its lexicon to suggest, this is reported as an *unknown word*. Interestingly, such flags are considered wrong in 24% of the occurrences on which the user performed any action, which above all only pertained to 2% of the overall unknown word flags. This means that if Acrolinx finds spelling suggestions for a word and creates a *general spelling flag*, it is probably a true spelling error. If it does not find such suggestions, however, the flagged *unknown word* is probably legitimate. This misclassification might be caused by special expert terminology being used in support forums which is missing in the Acrolinx terminology and spelling lexicon and which looks unlike anything Acrolinx knows. Therefore, *unknown word* flags should probably be disabled, too.

Rule	Type	Flag count	Flag attention	Flag precision	Suggestion proposal	Suggestion precision
a vs à	grammar	209	13%	89%	100%	100%
accord phrase nominale	grammar	189	12%	91%	90%	100%
accord sujet/verbe	grammar	124	15%	94%	94%	100%
ajoutez tiret	grammar	49	37%	89%	88%	100%
ajoutez un trait d'union	grammar	44	18%	88%	57%	100%
ajoutez une virgule	grammar	99	25%	92%	96%	100%
ce vs se	grammar	34	41%	100%	100%	100%
espace en trop	grammar	113	14%	100%	75%	100%
espaces autour ponctuation	grammar	602	25%	95%	86%	100%
forme verbale incorrecte	grammar	174	15%	96%	100%	100%
homophones verbe/nom	grammar	18	44%	100%	100%	100%
ça vs sa	grammar	16	75%	100%	100%	100%
élidez ce mot	grammar	139	30%	100%	86%	100%
évitez ponctuation	grammar	140	4%	83%	0%	N/A
sentence too long	style	247	2%	0%	0%	N/A
ajoutez virgule	style	21	33%	100%	100%	100%
ajoutez un blanc	style	17	47%	75%	100%	100%
erreur de majuscule	style	166	17%	86%	100%	100%
évitez adverbes	style	30	23%	71%	100%	100%
évitez le participe présent	style	52	15%	38%	100%	100%
évitez toute une phrase en majuscule	style	23	30%	86%	100%	100%
négation incomplète	style	118	31%	97%	100%	100%
ponctuation double	style	91	12%	91%	70%	100%
évitez abrév. forum	style	141	24%	94%	100%	100%
évitez le langage familier	style	53	15%	63%	100%	100%
évitez les anglicismes	style	39	15%	33%	100%	100%
évitez une conjonction en début de phrase	style	51	25%	46%	100%	100%
general spelling flag	spelling	1621	31%	89%	96%	97%
utilisez une majuscule en début de phrase	spelling	643	37%	97%	96%	99%

Table 25: Evaluation results for French pre-editing rules

Finally, the table shows that *if* the user considers a flag correct, Acrolinx provides a list of suggestions in the majority of cases, and those lists almost always include the correct suggestion. The only exception is represented by the rules "incorrect extra comma" and "sentence too long" as well as *unknown word* flags, which never have a suggestion. However, the latter two should be disabled anyway for reasons discussed above.

Overall, the results are quite encouraging, and show that, at least for those users who used the pre-editing functionality on a long-term basis, it proved to be quite valuable.

5.4 Results for French Pre-editing Rules

Table 25 shows the results for the French pre-editing rules. For French rules, the overall flag attention is higher than for English rules, though still relatively low. Again, this could be due to the

fact that many users may have just tried out the feature once. Also, the flag precision is better – 24 out of 29 rules have a precision of over 65%, 22 are even above 80%.

Overall, the findings are very similar, and similar conclusions can be drawn. In particular, the worst-performing rule is “sentence too long”. Other rules with a precision below average are “évitez une conjonction en début de phrase”, “évitez le langage familier”, and “évitez les anglicismes”, which all urge the user to change the type of language that is common in support forums. French forum users are apparently not keen to remove the colloquial tone either. Overall, the findings show that the pre-editing functionality was well received by users.

6. Conclusion

In this deliverable, we described the work performed in the final year of the ACCEPT project in order to evaluate the ACCEPT pre-editing and post-editing rules, as well as user interaction with rules. In addition to the large-scale human comparative evaluation carried out for the English-German language pair, we conducted smaller, focused experiments on the impact of pre-editing, as follows: analysis of degraded sentences (for German); study of the correlation between human and automatic evaluation results at the post level (for English-German) and at the sentence level (for French-English); and several studies on the impact of pre-editing across different data domains (for French-English). We investigated the impact of pre-editing across MT paradigms in a number of experiments, on a rule-by-rule basis as well as on a cumulative basis, both at the forum and at the sentence level, using versions of the Lucy and Systran MT systems specialised to the ACCEPT data domain. Together with insights from the analysis of user interaction with pre-editing rules, the results of these experiments confirm the high impact of rules, and specifically, the significant positive improvements in translation quality achieved in various settings. However, the impact of pre-editing on translation quality is generally not reflected by the results of automatic evaluation. With the exception of the English-German language pair, for which a statistically significant difference was observed in the scores of the METEOR metric, the automatic scores were not significantly different when the source text was pre-edited. The correlation between automatic and human evaluation results is positive but weak, the best correlation being achieved by METEOR for French-English at the post level; by BLEU at the sentence level; and by TER for English-German at the post level. These findings suggests that the capability of metrics to mirror human judgements depends largely on the experimental context.

The post-editing rules have been the object of evaluation in a dedicated experiment that focused on two dimensions of the MT output, namely, readability and usefulness for subsequent human post-editing. The results showed an overwhelmingly positive impact of automatic post-editing rules for French along both dimensions, confirmed by findings on increased similarity of output to reference translation, as computed with the TER metric.

Taken together, the evaluation results showed that the lightweight pre-editing and post-editing rules developed in the ACCEPT project for English and French user-generated content is highly useful for improving the quality of machine translation output, in both an SMT and RBMT setting, and for both in-domain and out-of-domain data.

References

- Byrt Ted, Janet Bishop, John B. Carlin:
  Bias, prevalence and kappa.
  Journal of Clinical Epidemiology, 46(5):423–429, 1993.
- Bontcheva, Kalina, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard, Niraj Aswani:
  TwitIE: An open-source information extraction pipeline for microblog text.
  In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2013)*, pages 83–90, Hissar, Bulgaria, September 2013.
- Pierrette Bouillon, Liliana Gaspar, Johanna Gerlach, Victoria Porro, Johann Roturier:
  Pre-editing by Forum Users: a Case Study
  In *Proceedings of the Workshop on Controlled Natural Language (CNL) Simplifying Language Use (CNL)*, pages 3−10, Reykjavik, Iceland, May 2014.
- Bredenkamp, Andrews, Berthold Crysmann, Mirela Petrea:
  Looking for errors: A declarative formalism for resource-adaptive language checking.
  In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, May-June 2000.
- Joseph L. Fleiss:
  Measuring nominal scale agreement among many raters.
  Psychological Bulletin, 76:378–382, 1981.
- Gerlach, Johanna, Victoria Porro, Pierrette Bouillon, Sabine Lehmann:
  La préédition 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)*, pages 539−546, Sables d'Olonne, France, June 2013.
- Gerlach, Johanna:
  Improving Statistical Machine Translation of Informal Language: a Rule-based Pre-editing Approach for French Forums.
  PhD thesis, University of Geneva, 2015.
- Kevin A. Hallgren:
  Computing inter-rater reliability for observational data: An overview and tutorial.
  Tutorials in quantitative methods for psychology 8(1), 23−34, 2012.
- Porro, Victoria, Johanna Gerlach, Pierrette Bouillon, Violeta Seretan:
  Rule-based automatic post-processing of SMT output to reduce human post-editing effort.
  In *Proceedings of the Translating and the Computer Conference*, London, U.K., November 2014.
- Roturier, Johann, Linda Mitchell, Robert Grabowski, Melanie Siegel:
  Using automatic machine translation metrics to analyze the impact of source reformulations.
  In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego, USA, October 2012.
- Seretan, Violeta, Pierrette Bouillon, Johanna Gerlach:
  A large-scale evaluation of pre editing strategies for improving user-generated content translation.
  In *Proceedings of the 9th Edition of the Language Resources and Evaluation Conference*, pages 1793-1799, Reykjavik, Iceland, May 2014.