Rule-based Automatic Post-processing of SMT Output to Reduce Human Post-editing Effort

Victoria Porro, Johanna Gerlach, Pierrette Bouillon, Violeta Seretan

Université de Genève FTI/TIM
40 Bvd. Du Pont-d'Arve, CH-1211 Genève 4, Suisse
{Victoria.Porro, Johanna.Gerlach, Pierrette.Bouillon, Violeta.Seretan}@unige.ch

Abstract

To enhance sharing of knowledge across the language barrier, the ACCEPT project focuses on improving machine translation of user-generated content by investigating pre- and post-editing strategies. Within this context, we have developed automatic monolingual post-editing rules for French, aimed at correcting frequent errors automatically. The rules were developed using the Acrolinx^{IQ} technology, which relies on shallow linguistic analysis. In this paper, we present an evaluation of these rules, considering their impact on the readability of MT output and their usefulness for subsequent manual post-editing. Results show that the readability of a high proportion of the data is indeed improved when automatic post-editing rules are applied. Their usefulness is confirmed by the fact that a large share of the edits brought about by the rules are in fact kept by human post-editors. Moreover, results reveal that edits which improve readability are not necessarily the same as those preserved by post-editors in the final output, hence the importance of considering both readability and post-editing effort in the evaluation of post-editing strategies.

Keywords: post-editing, statistical machine translation, user-generated content, language communities

1. Introduction

Since the emergence of the Web 2.0 paradigm, user-generated content (UGC) represents a large share of the informative content available nowadays. Online communities share technical information and exchange solutions to technical issues through forums and blogs. However, the uneven quality of UGC can hinder both readability and machine-translatability, thus preventing sharing of knowledge between language communities (Jiang et al., 2012; Roturier and Bensadoun, 2011).

The ACCEPT project [1] aims to improve the Statistical Machine Translation (SMT) of community content through minimally-intrusive pre-editing techniques, SMT improvement methods and post-editing strategies. The project targets two specific data domains: the technical forum domain, represented by posts in the Norton Community forum, and the medical domain, illustrated by Translators without Borders documents written by health professionals.

During the first year of the project, we found that pre-editing forum data significantly improves MT output quality (Lehmann et al., 2012; Gerlach et al., 2013a). Further work (Gerlach et al., 2013b) has shown that pre-editing which improves SMT output quality also has a positive impact on bilingual post-editing time. We are now developing post-editing rules intended to reduce post-editing effort, by automatically correcting the most frequent errors before submitting MT output to the post-editor.

This study focuses on the evaluation of the post-editing rules developed for French, and more specifically, on automatic rules designed for monolingual application. In the related literature,

[1] http://www.accept-project.eu/

there are several studies describing post-editing rules and evaluating them using automatic metrics or fluency-adequacy measures (Guzman, 2008; Valotkaite et al., 2012). However, to our knowledge, few such studies look into the actual use of the modifications produced by rules. We will assess: (1) the impact of the rules on the *readability* of the MT output and (2) their *usefulness* during the subsequent manual post-editing phase.

Our study relies on the following hypotheses: (1) the changes produced by our automatic monolingual rules contribute to making the text more readable; (2) automatic post-editing produces useful changes for the post-editing task and reduces technical effort; and (3) readability and usefulness for post-editing do not necessarily go hand in hand.

The paper is organised as follows. In Section 2, we show how post-editing research is performed in ACCEPT and describe the rules developed for French. In Section 3, we describe the experimental setup and provide details about data, tasks and participants. The results are analysed in Section 4, and conclusions and future work are presented in Section 5.

2. Post-editing in ACCEPT

In the ACCEPT project, post-editing rules, as well as pre-editing rules, are developed using the technology developed by one of our project partners, i.e., the Acrolinx^{IQ} engine (Bredenkamp et al, 2000). This rule-based engine uses a combination of shallow NLP components enabling the development of declarative rules, written in a formalism similar to regular expressions, based on the syntactic tagging of the text. A sample rule is displayed in Figure 1.

```
TRIGGER(80) == [@ne]? @auxFin^1 [@adv]* @verbInf^2
    -> ($aux, $inf)
    -> {mark: $aux, $inf;}
```

Figure 1. Rule formalism in Acrolinx^{IQ}

Rules can be applied through the ACCEPT portal interface (Seretan et al., 2014) or directly in any forum interface, using specific plugins that allow to check compliance with the rules (ACCEPT D5.6; Roturier et al., 2013).

The ACCEPT partners have so far explored several approaches to post-editing: manual vs automatic, monolingual vs bilingual (ACCEPT D7.2 and D2.4; Mitchell et al., 2013). For French texts machine-translated from English, we have focused on automatic monolingual rules for various reasons. Surface errors abound in machine-translated French texts. These errors seem a good target for source-independent lightweight rules that can be developed with simple patterns and shallow linguistic analysis. The automatic application of rules is motivated by two potential use scenarios. In a technical forum, where users have varied linguistic knowledge and might not have particular interest in fixing linguistic issues, automatic rule application requiring no participation or effort is clearly valuable. In a case where forum posts were to attain a better quality and a manual post-editing phase performed by bilinguals was necessary, automatically applying our rules beforehand could reduce both effort and time involved in this task.

We have developed 27 monolingual post-editing rules for French. The rules treat two types of phenomena: (1) spelling and grammar errors and (2) system-specific errors. We have used

different resources to develop and infer the rules: manual analysis of previously post-edited data, bilingual terminology extraction on source and raw translation, and spell-checking of the raw translation using Acrolinx[IQ].

Examples of errors and monolingual automatic rules for French can be found in Table 1.

Incorrect negation *Je ~~n'~~ai accès à distance.* *~~C'est~~ Ce n'est pas bloqué par le fichier.*	**Wrong word order** *Le ~~Norton~~ technicien Norton m'a conseillé de [...].* *Votre ~~PC~~ périphérique PC doit être [...].*
Incorrect punctuation and elision (comma, hyphen) *Je ~~comprends mais~~ comprends, mais...* *~~As-tu~~ As-tu lu ça ?* *Est-ce ~~qui il~~ qu'il s'agit de...* *Blocage des appels Pas de message > appels. Pas de*	**Reformulation** *~~Je suis en espérant~~ > J'espère* *~~Veuillez aider.~~ > Aidez-moi, s'il vous plaît.* *~~Hi~~ Bonjour, merci pour le message.* *J'espère que ~~cette~~ ça aide.*
Incorrect verb form (imperative, infinitive, participe, subjonctif) *Il n'a pas ~~faire~~ fait ça.* *J'ai dû ~~fait~~ faire ça.* *Bien que je ne ~~comprends~~ comprenne ça, [...].* *~~Regardes~~ Regarde en bas.*	**Agreement errors (subject-verb, determinant-noun, noun-adjective)** *Lorsque je ~~faire~~ fais une recherche [...].* *commentaires ~~apprécié~~ appréciés* *~~nouveau~~ nouvel article* *~~le~~ les deux ~~chose~~ choses*
Casing error *Il a demandé à ~~Si~~ si je savais [...].* *~~tout~~ Tout en supposant que [...].*	**Wrong term and anglicisms** *Les mises à jour ~~norton~~ Norton [...].* *Veuillez la mettre à jour ~~asap~~ au plus tôt.*
Missing or extra spaces *4GB > 4 GB* *Est-il bloqué? > Est-il bloqué ?*	**Doubled words** *Je ne l'ai ~~pas pas~~ pas fait.* *Re: Piratage ~~d'~~du navigateur.* *J'ai mis à jour ~~mon les~~ mes pilotes.*
Avoid direct questions *~~Tu as~~ As-tu lu le message ?*	

Table 1. Example of phenomena treated by French automatic post-editing rules

3. Experimental Setup

In this section, we describe the methodology followed to test our hypotheses. We introduce the tasks designed to this end, the data selected and the participants recruited for the study.

3.1 Method

In our study, two tasks were designed to evaluate automatic monolingual post-editing rules in terms of readability and usefulness (as discussed in Section 1): a comparative evaluation task aimed at eliciting judgments on the impact of our rules on *readability* (the extent to which a translated segment reads naturally), and a post-editing task aimed at determining the *usefulness* of changes introduced by rules in an actual post-editing context (their beneficial and practical use). Results for readability and usefulness were cross-analysed.

3.2 Data

An original corpus of 5000 English sentences extracted from the Norton Community forum was pre-edited using the project's pre-editing rules for English. The data was then translated

into French using the project's baseline system, which is a phrase-based Moses system, trained on translation memory data supplied by our partner, Symantec, and supplemented with Europarl and news-commentary data (ACCEPT D4.1).

We automatically applied our post-editing rules to the translated corpus and removed sentences with more than 40 words to avoid long sentences. We classified the resulting sentences according to the Levenshtein distance between the automatically post-edited (APE) output and the raw output, and then according to the number of rules that had been applied in each APE sentence. Our intention was to focus on sentences with the highest number of changes in order to cover a larger number of post-editing rules. We kept for this study a sample consisting of the first 200 sentences appearing at the top of the resulting classification. One sentence was duplicated and eliminated from the selection. The selected 199 sentences totalled about 3700 words.

3.3 Participants

For both tasks performed in this study, we recruited three translation students in the second year of the MA programme at the Faculty of Translation and Interpreting of the University of Geneva. They are native French speakers with English as their main working language. None of the participants had specific technical knowledge.

3.4 Comparative Evaluation Task

This task was meant to test our first hypothesis (see Section 1). We let annotators comparatively evaluate pairs of raw and APE sentences. They rated each pair on a 3-point scale: first better–equal–second better, according to which of the versions they considered to be more readable. For this task, annotators were not shown the corresponding source. The evaluation focussed on readability alone, with no consideration of adequacy. The two versions were shown to annotators in random order to avoid bias. In addition to evaluating the overall readability of the sentence, the annotators rated all individual edits (IE)[2] automatically introduced by our rules using the same 3-point scale mentioned above. Annotators were provided with guidelines and evaluated 199 sentences and 391 IEs using Excel sheets.

3.5 Post-editing Task

To test the second hypothesis, we asked the same annotators to manually post-edit the APE output with access to the source text.

The post-editing task was performed using the post-editing environment of the ACCEPT portal (ACCEPT D5.6; Roturier et al., 2013) in bilingual mode, as shown in Figure 2. Participants were provided with post-editing guidelines and a glossary of the domain. They were asked to render a grammatically correct target sentence, which should convey the same meaning as the original, and to use as much of the raw MT output as possible. Style was not given priority. No time limit was given, and all participants were paid.

[2] We understand by "individual edits" any sequence of adjacent words modified by the automatic application of our monolingual rules for French.

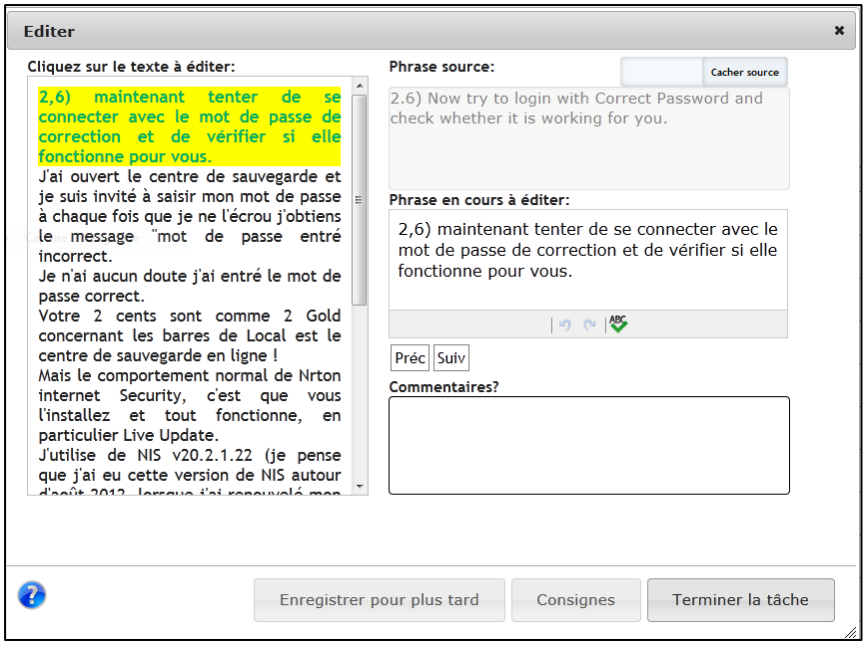

Figure 2. Interface of the ACCEPT post-editing environment

Once the task was completed, we compared the resulting sentences with the APE version to identify the actual differences between the versions. We used an in-house tool to automatically identify the IEs that our rules had introduced in the raw output and that human post-editors had kept during the subsequent manual processing. This allowed us to check the rate of preservation of the IEs. The results of the automatic extraction were checked manually to ensure that all IEs were detected, including insertion/deletion of spaces and use of capitals.

4. Results

This section presents the results obtained by applying the method described in Section 3. We proceed by presenting the findings related to the hypotheses put forward in Section 1.

4.1 Comparative Evaluation Task - *Readability*

The number of sentences and individual edits (IEs) deemed as better was higher in the case of automatically post-edited (APE) sentences than in the case of raw sentences. A total of 199 sentences and 391 IE were evaluated. The results of the comparative evaluation task are shown in Table 2.

Ratings were similar both at the sentence level and at the IE level. On average, 74% of sentences (78%-80%-64%) and 75% of IE (74%-84%-68%) were considered better in terms of readability when automatic post-editing is applied. While in each evaluation a mean of 20% of annotated pairs were considered equal, the amount of raw sentences and IEs considered better was negligible (1% to 6%). A one-sample chi-square test of goodness of fit was performed to test the difference in proportions. For all three annotators, the difference in proportions between the three categories is significant at sentence level ($\chi^2(2, N=199) = 186.5 / 211.4 / 114, p < 0.001$) and at IE level ($\chi^2(2, N=391) = 298.6 / 455.5 / 255.7, p < 0.001$).

	Sentence level		
	APE better	***equal***	***raw better***
Annotator 1	156 (78%)	34 (17%)	9 (5%)
Annotator 2	161 (80%)	36 (18%)	2 (1%)
Annotator 3	128 (64%)	66 (33%)	5 (3%)
	IE level		
Annotator 1	288 (74%)	80 (20%)	23 (6%)
Annotator 2	328 (84%)	51 (13%)	12 (3%)
Annotator 3	268 (69%)	111 (28%)	12 (3%)

Table 2. Comparative evaluation task − Results for readability

The observed agreement for judgements at the IE level was of 56% (unanimous = 219/391) and it reached 58% at the sentence level (unanimous = 115/199).

We assessed inter-annotator agreement (IAA) to validate this observation. At the sentence level, we first calculated Cohen's kappa for each pair of annotators (Cohen, 1960). Although the observed agreement was relatively high, results only showed *fair* agreement (average $k = 0.277$), probably due to the effects of prevalence (Artstein&Poesio, 2008). Because k may become unreliable when used on skewed data, we decided to assess IAA using a two-way intra-class correlation (ICC) (McGraw, 1996). The resulting ICC was in the *good* range, ICC = 0.64 (Cicchetti, 1994), indicating that annotators had a relatively high degree of agreement and a low amount of measurement error.

We did the same for the evaluation at the IE level. Cohen's kappa was equally low (average $k = 0.245$) and the two-way ICC (McGraw, 1996) was also in the *good* range, ICC = 0.62 (Cicchetti, 1994).

The results of this first experiment confirm our hypothesis that our automatic monolingual rules significantly improve readability.

4.2 Post-editing Task - *Usefulness*

For this task, the analysis focused on the IE level. We assessed the rate of preservation of the 391 IEs that had been introduced by our automatic rules.

Our analysis showed that a high percentage of IEs (70%) was kept during manual post-editing, suggesting that our rules perform useful modifications that reduce the number of changes post-editors have to perform to reach the final output (see Table 3). Some sentences were not edited at all (4%, 8%, 10%). A one-sample chi-square test of goodness of fit was performed to test the difference in proportions between the Found and Missing at IE level. For all three annotators, the difference is significant ($\chi^2(1,\ N{=}391) = 64.7 / 68 / 78.3,\ p < 0.001$). Results are shown in Table 3.

	IE level		Sentences
	Found	***Missing***	***No edits***
Annotator 1	275 (70%)	116 (30%)	8 (4%)
Annotator 2	277 (71%)	114 (29%)	16 (8%)
Annotator 3	283 (72%)	108 (28%)	21 (10%)

Table 3. Post-editing task − Results for usefulness

To assess agreement of the three post-editors on the IEs that were kept, we again computed Cohen's kappa (Cohen, 1960), which showed moderate agreement, $k = 0.559$ (Landis&Koch, 1977). A two-way ICC assessment indicated *excellent* agreement, ICC = 0.79 (Cicchetti, 1994).

To quantify the share of work performed by the automatic post-editing rules, we chose to measure edit distance by means of TER (Snover et al., 2006). Our assumption was that the TER score would be lower for the automatically post-edited (APE) output than for the raw output.

We computed TER for the raw MT and APE output using the manually post-edited sentences as reference. The raw MT output achieved a TER score of 0.42, while the APE output dropped to 0.27. This suggests that, in terms of edits, our rules contribute to making the MT output more similar to the human output (lower values indicating higher similarity).

Since the manual post-editing was done by using the automatically post-edited as a basis, it might be argued that the final human output will be closer to the APE version than to the raw MT output because of this methodological choice. To obtain scores not subject to this bias, we computed TER scores against a human reference built from scratch. This reference was produced by a native French speaking professional translator with domain knowledge. The translator used the same guidelines as the post-editors. Against this second reference, the raw MT output achieved a TER score of 0.66 against 0.59 for the APE version. While the difference between scores is smaller, it is still in favour of the APE version. These results confirm that the changes introduced by the automatic rules bring the text closer to the final version and reduce the post-editing "technical effort" (as defined by Krings, 2001).

In view of the above, we can conclude that our second hypothesis was also confirmed. We had assumed that most individual edits (IE) introduced by our rules would be kept in the final version of the selected sentences.

4.3 Readability vs Usefulness

Our third hypothesis was that the IEs preserved during manual post-editing would not necessarily be the same as those IEs judged as enhancing readability. We expected a low correlation between readability and usefulness. To test the hypothesis, we crossed the data obtained in the comparative evaluation task (*readability*) and the post-editing task (*usefulness*).

The cross-data analysis was very similar for all three annotators (see Table 4). Results show that, on average, 60% of the IEs introduced by our rules (58%-66%-54%) were considered *better* during the comparative evaluation task and also kept during the manual post-editing of the output. A lower percentage (16%-17%-16%) was discarded. The rate of preservation of IEs considered *equal* was of about 10%, while a similar percentage (11%) was discarded. Finally, only 1% to 3% of the raw versions were found better and either discarded or kept.

		IE level		
		Annotator 1	**Annotator 2**	**Annotator 3**
APE better	***Found***	227 (58%)	260 (66%)	213 (54%)
APE better	***Missing***	61 (16%)	68 (17%)	55 (16%)
Equal	***Found***	38 (10%)	15 (4%)	64 (16%)
Equal	***Missing***	42 (11%)	36 (9%)	47 (12%)
Raw better	***Found***	10 (3%)	2 (1%)	6 (2%)
Raw better	***Missing***	13 (3%)	10 (3%)	6 (2%)

Table 4. Combined results for readability and usefulness at the IE level

To assess the correlation between *readability* and *usefulness*, we calculated Kendall's tau (Kendall, 1938). Results showed a weak positive correlation, $\tau = 0.306$ (Evans, 1996), statistically significant ($p < 0.01$). This weak correlation between *readability* and *usefulness* is unsurprising and confirms our hypothesis.

Table 5 illustrates correlation cases. We provide one example for each case, but for space limitations, we will only comment on APE better-Found and APE better-Missing cases. A thorough study of all correlation cases is still needed to draw complete and definitive conclusions.

	English Source	**Raw MT-Output**	**APE sentence**
R: **APE better** U: **IE Found**	Also are there any programs you recommend doing the job?	Il y a également des programmes que vous recommande de faire ce travail?	Y a-t-il également des programmes que vous recommande de faire ce travail ?
R: **APE better** U: **IE Missing**	If you have already done this, but the space is not showing released [...].	Si vous avez déjà fait, mais l'espace n'est pas faire preuve de la sortie de message privé [...].	Si vous avez déjà fait, mais l'espace n'a pas fait preuve de la sortie de message privé [...].
R: **Equal** U: **IE Found**	Are they still valid?	Ils sont toujours valables?	Sont-ils toujours valables ?
R: **Equal** U: **IE Missing**	I was looking for Ghost [...]	J'étais en train d'pour Ghost [...]	J'étais en train de pour Ghost [...]
R: **Raw better** U: **IE Found**	Is the post below also posted by you?	C'est le post ci-dessous également publiés par vous?	Est-ce le post ci-dessous également publiés par vous ?
R: **Raw better** U: **IE Missing**	Norton employees have their names in bold red letters.	Norton les employés ont leurs noms en gras lettres rouges.	Les employés Norton ont leurs noms en grasses lettres rouges.

Table 5. Sample cross-data for *readability* and *usefulness*

Combinations APE better-Found are the most common. They correspond mostly to corrections of shallow errors related to grammar and structure, and some specific reformulations. Row one of Table 5 illustrates this correlation.

Combinations APE better-Missing are the second most common. These can be explained by the characteristics of the rules themselves. Due to the chosen technology and to the fact that monolingual rules do not refer to the source text, our rules are incapable of correcting long distance dependencies, detecting incorrect lexical choices, producing perfect agreement between words or choosing the right verb tenses. They are developed to treat mainly local and highly recurring phenomena. As a consequence, a rule might correct a sequence of words and thereby locally improve the natural flow of the text (e.g., by inversing verb and subject in questions or correcting wrong verb forms), but when considering the entire sentence, these changes might not be relevant. Row two of Table 5 illustrates these cases. The APE version was considered better in the readability task, but the individual edits introduced were not kept. Taken out of context, the sequence "*n'a pas fait preuve de*" is better than "*n'est pas faire preuve de*". However, considering the entire sentence and the source text, both versions are wrong and the correction made by the IE is useless.

5. Conclusions

Our study has shown that lightweight automatic post-editing rules such as the ones developed in the ACCEPT project for French user-generated content are beneficial both in terms of *readability* and *usefulness* for subsequent manual post-editing. About 74% of the sentences and IEs evaluated were deemed better when automatic post-editing rules were applied, and 70% of the IEs that the rules had introduced were kept. The TER results confirm that an APE version can reduce post-editors' technical effort.

The cross-data analysis confirmed that certain rules induce changes that are more adequate for readability purposes than for the actual post-editing task. This analysis has allowed us to better understand and explain why our rules may produce a divergent effect, that is, they improve readability but do not help in the post-editing task, or vice versa. Although a high percentage of IEs improve readability and are useful for manual post-editing, a non-negligible percentage fell in other categories.

In future work, we plan to perform a more detailed analysis of the results obtained in this study. In particular, we want to look into the specific rules that produce the divergent effect mentioned above. This will allow us to classify and filter rules depending on the purpose they may serve the best. We also plan to perform the extrinsic evaluation of post-editing rules, in an actual forum context. Since the rules are tailored to social platforms and in particular to technical forums, we would like to perform evaluation using real users, in order to assess both the readability of rules and their contribution to solving users' problem at hand.

Acknowledgements

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.

References

ACCEPT Deliverable 2.4 (2014). Available at: http://www.accept.unige.ch/Products/D-2-4-Definition-of-Post-editing-Rules.pdf

ACCEPT Deliverable 4.1 (2012). Available at: http://www.accept.unige.ch/Products/D_4_1_Baseline_MT_systems.pdf

ACCEPT Deliverable 5.6 (2013). Available at: http://www.accept.unige.ch/Products/D_5_6_Browser-based_client_demonstrator_and_adapted_post-editing_environment_and_evaluation_portal_prototypes.pdf

ACCEPT Deliverable 7.2 (2013). Available at: http://www.accept.unige.ch/Products/D_7_2_Report_on_assistance.pdf

Artstein, R & Poesio, M (2008). Inter-coder agreement for computational linguistics, *Comput. Linguist.*, 34, pp. 555–596.

Bredenkamp, A, Crysmann B & Petrea, M (2000). Looking for errors: A declarative formalism for resource-adaptive language checking. In *Proceedings of LREC 2000*, Athens, Greece.

Cicchetti, DV (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology, *Psychological Assessment*, 6, p. 284.

Cohen, J (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20(1), pp. 37–46.

Evans, JD (1996). *Straightforward statistics for the behavioral sciences*, Pacific Grove, CA: Brooks/Cole Publishing.

Gerlach, J, Porro, V, Bouillon, P, & Lehmann, S (2013a). La préédition avec des règles peu coûteuses, utile pour la TA statistique des forums ? In *Proceedings of TALN/RECITAL 2013*, Sables d'Olonne, France.

Gerlach, J, Porro, V, Bouillon, P, & Lehmann, S (2013b). 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.

Guzmán, R (2008). Advanced automatic MT post-editing, *Multilingual*, vol.19, issue 3, pp. 52–57.

Jiang, J, Way, A, and Haque, R (2012). Translating user-generated content in the social networking space. In *Proceedings of AMTA 2012*, San Diego, USA.

Kendall, M (1938). A new measure of rank correlation, *Biometrica*, 30(1–2), pp. 81–89.

Krings, HP (2001). *Repairing texts: Empirical investigations of machine translation post-editing process*, The Kent State University Press, Kent, OH.

Lehmann, S, Gottesman, B, Grabowski, R, Kudo, M, Lo SKP, Siegel, M & Fouvry F (2012). Applying CNL authoring support to improve machine translation of forum data. In Kuhn, T., Fuchs, N. (eds) *Controlled Natural Language. Third International Workshop,* pp. 1–10.

McGraw, KO & Wong, SP (1996). Forming inferences about some intraclass correlation coefficients, *Psychological Methods*, 1, pp. 30–46.

Mitchell, L, Roturier, J & O'Brien, S (2013). Community-based post-editing of machine-translated content: monolingual vs. bilingual. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, Nice, France.

Roturier, J, & Bensadoun, A (2011). Evaluation of MT systems to translate user generated content. In *Proceedings of the MT Summit XIII*, pp. 244–251.

Roturier, J, Mitchell, L & Silva, D (2013). The ACCEPT post-editing environment: A flexible and customisable online tool to perform and analyse machine translation post-editing. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, Nice, France.

Seretan, V, Roturier, J, Silva, D, & Bouillon, P (2014). The ACCEPT Portal: An online framework for the pre-editing and post-editing of user-generated content. In *Proceedings of HaCaT*, Gothenburg, Sweden.

Snover, M, Dorr B, Schwartz, R, Micciulla & L and Makhoul, J (2006). A sudy of Translation Edit Rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas*, Cambridge, Massachusetts.

Valotkaite, J & Asadullah, M (2012). Error detection for post-editing rule-based machine translation. In *Proceedings of the AMTA 2012-WPTP2*, San Diego, USA.