Community-based post-editing of machine-translated content: monolingual vs. bilingual

Linda Mitchell[†], **Johann Roturier**[†], **Sharon O'Brien**[‡]
[†] Symantec Ltd., Ballycoolin Business Park, Blanchardstown, Dublin 15, Ireland
{linda_mitchell, johann_roturier}@symantec.com
[‡] School of Applied Languages and Intercultural Studies, Dublin City University, Ireland
sharon.obrien@mail.dcu.ie

Abstract

We carried out a machine-translation post-editing pilot study with users of an IT support forum community. For both language pairs (English to German, English to French), 4 native speakers for each language were recruited. They performed monolingual and bilingual post-editing tasks on machine-translated forum content. The post-edited content was evaluated using human evaluation (fluency, comprehensibility, fidelity). We found that monolingual post-editing can lead to improved fluency and comprehensibility scores similar to those achieved through bilingual post-editing, while we found that fidelity improved considerably more for the bilingual set-up. Furthermore, the performance across post-editors varied greatly and it was found that some post-editors are able to produce better quality in a monolingual set-up than others.

1 Introduction

User-generated content, such as in the Norton support forums [1], which provide a platform for solving problems related to Norton products online in several languages, is growing rapidly. It is only useful to those, however, who have sufficient knowledge of the language it was composed in. To broaden the impact of this content and to provide solutions to users faster, a combination of machine translation and post-editing is explored as an option. Rather than having translation professionals perform post-editing, opening it up to users of the community, who are domain experts, goes hand in hand with the concept of users supporting users. The research reported here does not investigate community users' notions of adequate quality, but quality assessment of community post-editing will be a focus of a future, extended study. The focus of post-editing research to date has been primarily on professional translators. It has been noted that translators' attitudes towards post-editing can be problematic, that there is considerable individual variation among post-editors and that experienced translators tend to ignore post-editing guidelines (de Almeida and O'Brien 2010). This raises the question of whether groups other than professional translators might be able to perform post-editing successfully. One idea that has been suggested recently is that post-editing might be done adequately by monolingual users (Koehn 2010), which is the focus of our pilot study. German and French native speakers, users of the Norton communities were recruited via private message and public announcement to post-edit machine translated content in a monolingual and a bilingual environment. Thus, this study focusses on community-based post-editing, involving a community that is already existent and has a main purpose other than translation/post-editing, here IT support. This has to be distinguished from crowd post-editing, which involves a community of users whose main purpose it is to translate or post-edit. This study focusses on comparing the two set-ups, rather than the two language pairs. The novel contributions of the paper are as follows: 1) Evaluating the MT post-editing output provided by community members; 2) Comparing monolingual and bilingual post-editing performance for User Gen-

[1] http://community.norton.com/

erated Content; 3) Identifying characteristics that make monolingual post-editing difficult.

2 Related Work

Post-editing has received attention increasingly over the last years (e.g. Guerberof 2009, Garcia 2010, Koponen 2010). Bilingual post-editing has been the main focus so far, for the obvious reason that it is assumed that bilingual competence is a pre-requisite for successful post-editing. However, there have been studies tackling monolingual post-editing (e.g. Hu et al. 2010, Koehn 2010, Lin et al. 2010) with tentative positive results. Monolingual post-editing has also served as an interim step in the evaluation of machine translated content, as for example presented in the WMT09 data (Callison-Burch et al. 2009).

3 Experimental Set-Up

Due to restricted resources, the participants for this study were required to complete both monolingual and bilingual post-editing tasks[2], which also ensured comparability between those two set-ups. The aim was to get an overview of what kind of output community post-editors can produce in a bilingual and a monolingual set-up considering their knowledge of English and the Norton products and to identify types of segments that are difficult for community post-editors in order to be able to optimise the MT system and the post-editing process. Thus, four users were recruited for each language pair, with one participant (for EN-DE) completing monolingual tasks only[3] and the others completing both bi- and monolingual tasks.

The machine translation system used in this study[4] was trained on bilingual data both from in-domain data, e.g. product manuals of Norton products, and out-of-domain data, i.e. WMT12 releases of EUROPARL and news commentary (EN-DE, EN-FR) using Moses (Koehn et al., 2007). When training an SMT system, it is preferable to use a corpus that is close to the texts that will be translated with it (in-domain), i.e. in this context domain specific texts. Out-of-domain data was used as supplementary data to enrich and increase coverage of lexical resources. The test set was taken from the English-speaking support forum. They consist of the original question in a thread, its subject line and the post that had been marked as the solution to the question in the forum. The content to be post-edited was taken from a set of 347 texts[5], which had been extracted previously for the purpose of machine translation.

3.1 Clustering Technique

It was believed to skew the post-editing times if the participants were to edit each task more than once. Thus, a method of clustering similar posts together was deployed. Rather than selecting posts randomly and forming two groups, which may have resulted in two sets of posts that are quite different given the small number of posts selected, clustering ensured that the posts in both groups were as similar as possible in terms of characteristics described below. Characteristics considered in this clustering technique were meta statistics like text length (word count), sentence length, type-token-ratio (TTR), as well as content which is expressed in number of maskable tokens and perplexity with respect to a bigger forum-based language model (LM). The forum-based language model is a 5-gram LM with modified Kneser-Ney (Kneser and Ney, 1995) smoothing trained on the available monolingual English forum data (approx. a million sentences). It was trained using the IRSTLM (Federico et al., 2008) language modelling toolkit. To automatically achieve this, an unsupervised clustering approach based on the K-mean clustering approach (MacQueen, 1967) and more specifically the open source K-Means algorithm in the Weka Toolkit were used. The K-means clustering approach aims to group n observations into k groups to assign each observation to a group with the nearest mean. Four clusters were obtained out of which two tasks were were selected randomly from each of the clusters for the monolingual set-up and one task was selected randomly from each of the clusters for the bilingual set-up (in total: 8 monolingual tasks, 4 bilingual tasks).

Table 1 displays the number of segments for each set-up (monolingual and bilingual) and the number of words. The average number of seg-

[2]The English skills of the participants varied, i.e. the fact they were not bilinguals (cf. sections 4 and 5). The bilingual set-up merely indicates that they had access to the source text.

[3]This participant dropped out of the study after completing the monolingual tasks. This was beyond our control as the participants were volunteers.

[4]http://www.accept.unige.ch/Products/D_4_1_Baseline_MT_systems.pdf

[5]With each text containing a subject line, question and answer.

ments for each task was 8 and the average word count was 140 words.

Set-up	Tasks	Segments	Words
Monolingual DE	8	75	1125
Bilingual DE	4	28	504
Monolingual FR	8	70	1078
Bilingual FR	4	29	504

Table 1: Number of Tasks, Segments and Words per Set-up

3.2 Tasks

The users performed the post-editing tasks using a portal that was especially developed for post-editing, the interface of which is displayed in Figure 1. The interface offered the following functionality: undo/redo, spelling and grammar checking and access to alternative words for four of the monolingual tasks. The left half of the window shows the full text to be edited for that particular task. In the top right edit box the user can edit the current segment. Comments can be made in the edit box at the bottom right. All edits were saved automatically. During the post-editing process, editing time, keystrokes, usage of translation options etc. (cf. Roturier et al. 2013) were recorded in the portal. The following guidelines were displayed by clicking on the "Guidelines" button:

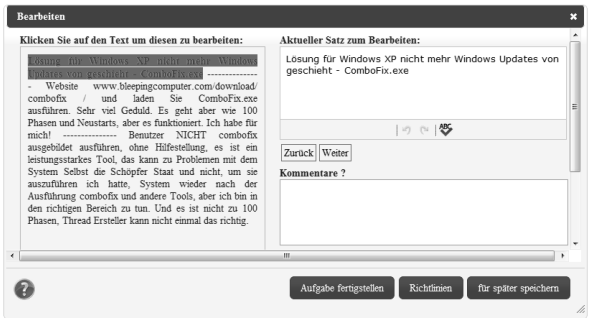

Figure 1: Post-Editing Interface

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

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

3.3 Evaluation

For human evaluation, three criteria were considered: fluency, comprehensibility and fidelity. The scales used for fluency and fidelity were taken from LDC (2002). The scale for comprehensibility was adopted from a previous study (Roturier and Bensadoun 2011). All three were measured on a 5-point Likert scale (0-4). The evaluation for this pilot study was carried out by two authors of this paper (one per language pair), native speakers of the target languages[6]. The segments for the MT output and the post-edited output were rated separately and the scores were then compared. The raw MT output and the post-edited output were also rated using the TER (Snover et al. 2006) automatic metrics, comparing them to two sets of reference translations, provided by a language service provider. One set using formal language and one set with a more informal style (i.e. colloquial language) were thus used for investigating whether the post-edited segments are closer to formal or informal language on the assumption that the language used in user-generated content would more closely approximate the informal reference language.

4 Results

Table 2 shows the scores (human evaluation) of the raw MT output compared to the post-edited content. For this, the scores from the human evaluation were added for all users for each task in the set-ups (monolingual and bilingual). They are broken down into percentages of all segments that were improved, that retained their score or that were diminished in their scores for the monolingual and the bilingual set-up. Fluency increased the most for both set-ups followed by comprehensibility and fidelity. The table shows that for monolingual post-editing PE performs better than

[6]The evaluation was not blind as one of the evaluators was in charge of the study.

the baseline MT system in terms of fluency for 67.3%, and in terms of comprehensibility for 57% of all segments. These figures are quite close to the scores for bilingual post-editing. For comprehensibility, the number of degradations stayed the same. Bilingual post-editing resulted in a higher number of improved segments for fidelity. What is striking, however, is that fidelity increased for 43% of the segments for monolingual post-editing. It should also be noted that there was a considerable percentage of degradations for fidelity in the monolingual set-up (28%) and the bilingual set-up (20%). The results of this pilot study suggest that the monolingual set-up leads to similar results in terms of improvements and degradations in fluency and comprehensibility compared to the bilingual set-up. It also leads to a greater number of improved segments for the bilingual set-up, with a considerable number of degradations, however.

	fluency %	**compr. %**	**fidelity %**
mono.			
improved	67.3	57	43
same	20.4	30	29
worse	12.3	13	28
bilingual			
improved	**70.2**	**64**	**56**
same	15.5	23	24
worse	14.3	13	20

Table 2: Human evaluation (German)

Table 3 shows the results for the French part of the experiment. There is little difference in the percentages between the two set-ups for fluency and fidelity. Comprehensibility scores the lowest, with the number of improved segments increasing remarkably for the bilingual set-up. This could be due to short post-editing times (cf. Figure 4. For the bilingual set-up, however, the scores for comprehensibility are considerably higher, which is also the biggest improvement of all (14 points). This suggests that the presentation of the English source text did make a difference in comprehensibility. It also needs to be considered that the number of improved scores for fidelity falls by three points and the number of degradations by four points from the monolingual set-up to the bilingual set-up. The present data suggests that for French there does not seem to be a great difference for fidelity across the two set-ups. A possible reason for this would be that the French post-editors had a better knowledge of the domain than the German ones or that English skills influenced the post-editing results less for the French participants than for the German participants. A study of a larger scale would be necessary to confirm these suggestions.

	fluency %	**compr. %**	**fidelity %**
mono.			
improved	**63**	48.6	**67**
same	20	25.5	18
worse	17	25.9	15
bilingual			
improved	**63**	**63**	64
same	27	26	25
worse	10	11	11

Table 3: Human evaluation (French)

4.1 Evaluation Per User - Summary

Table 4 shows the percentages of segments improved, that stayed the same and deteriorated for the German participants grouped by category (fluency etc.) with the best score marked in all categories. Self-reported knowledge of English and the Norton products was measured on a 5-point Likert scale (1-5) and is displayed along with the rank[7] of the post-editors according to the performance displayed in the top part of the table. For German, for all four participants it is true that the two skills combined, rather than just one of the two skills, correlate with the participants' ranks (their performance).

As displayed in Table 4, participant B had the biggest increase of improved segments for all three evaluation criteria. It is noteworthy that for French (Table 5) there is also one outstanding participant (B). For the French participants, the self-reported English skills and knowledge of the Norton products do not seem to correlate to their actual performance (rank). However, the values are very similar for both the skills and the percentages of improved segments. In order to draw a conclusion here, the skills would need to be tested to avoid bias and a larger number of participants would be needed.

Table 6 (German) and 7 (French) present the TER scores obtained by (i) comparing the MT output against the segments produced by each user

[7]The rank was calculated by adding the number of improvements for fluency, comprehensibility and fidelity for each participant and subtracting the number of degradations for the same.

[8]Participant D only completed monolingual tasks. Thus, the rank for D is based on those.

Participant:	A	B	C	D[8]
fluency in %				
improved	45	**77**	76	50
same	51	20	24	47
worse	4	3	0	3
comprehensibility in %				
improved	36	**70**	65	39
same	60	28	34	55
worse	4	2	1	6
fidelity in %				
improved	24	**53**	51	13
same	61	41	43	74
worse	15	6	6	13
rank (absolute)	3	1	2	4
skills (Likert 1-5)				
English knowledge	3	5	3	2
Norton knowledge	2	4	4	2

Table 4: Human Evaluation Across All Tasks Per Participant (German)

Participant:	A	B	C	D
fluency in %				
improved	54	**63**	51	57
same	39	29	34	40
worse	7	8	15	3
comprehensibility in %				
improved	34	**52**	42	45.5
same	60	32	34	45.5
worse	6	16	24	9
fidelity in %				
improved	57	**67**	53	59
same	38	25	27	40
worse	5	8	20	1
rank (absolute)	3	1	4	2
skills (Likert 1-5)				
English knowledge	3	3	3	4
Norton knowledge	4	3	4	4

Table 5: Human Evaluation Across All Tasks Per Participant (French)

and (ii) comparing the output of each user against the reference translations (regardless of the post-editing set-up) in the TER-1 and TER-2 columns. It was hoped to obtain some insight into whether the Translation Edit Rate can be used as an indicator of quality (in regards to human evaluation) here. The nature of the pilot study does not allow for computing statistical significance reliably. The trends presented thus need to be investigated further.

TER-1 refers to the reference translation set that was obtained with the instructions to use formal language and TER-2 to use informal language, in order to identify whether the MT output and the post-edited output are closer to formal or informal language. As can be seen in Tables 6 and 7 which contain TER scores comparing the MT segments with the post-edited segments, the TER scores are consistent with the percentages of improved segments in Tables 4 and 5 (across all categories). That means, the more the participants changed the MT output of a segment, the better the segments scored in terms of fidelity, comprehensibility and fluency. This is the case for all users, apart from for users A and C for French. When comparing the post-edited segments with the reference translations, however, the TER scores are not consistent with the percentages of improvements observed during human evaluation. While the best post-editor (based on ranking) for the German language pair (participant B) produces content that is the closest to the reference translations, the second best post-editor (participant C) produces content that differs most from the reference translations. It is not as clear for French, as the output of the best performing post-editor (participant B) is marginally closer to the reference translations compared to that of the other post-editors. Thus, comparing the post-edited output to the MT output appears to give some indication in regards to quality (as judged by humans for the criteria fluency, comprehensibility, fidelity), whereas the comparison of post-edited output to the reference translations does not.

[9]TER-1 and TER-2 refer to the two sets of reference translations. Both sets of values are calculated using TER.

	MT	Reference[9]	
User	TER	TER-1	TER-2
MT	N/A	72.2	66.9
A	32.3	75.1	70.6
B	66.4	71.3	68.9
C	47.7	75.4	71.8
D	32.9	73.8	71.0

Table 6: Automatic Metrics per Participant (German)

	MT	Reference	
User	TER	TER-1	TER-2
MT	N/A	79.1	73.3
A	20.5	77.2	73.2
B	46.9	76.8	73.1
C	29.3	77.9	73.4
D	39.8	77.4	73.2

Table 7: Automatic Metrics per Participant (French)

4.2 Monolingual vs. Bilingual

The high percentages of segments improved in terms of fluency for both French and German can

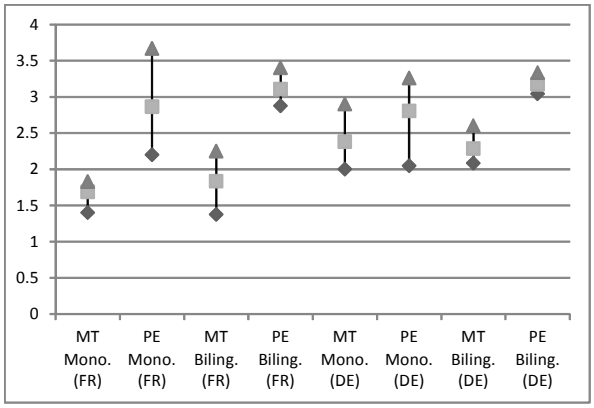

Figure 2: Fidelity Scores (Human Evaluation) with Minimum, Average and Maximum marked, with MT Mono (FR) meaning fidelity scores for the raw MT output intended for monolingual post-editing for French etc.

be attributed to the fact that the source text is not always needed to make a text fluent. Figure 2 displays the quality of the post-edited content in contrasting the range of fidelity scores (how much of the source content was retained) of the raw MT output with that of the monolingually post-edited content and the bilingually post-edited content for both French and German. It is evident that there is a wider variation in fidelity scores for the monolingual set-ups than for the bilingual set-ups. The reason for the highest percentage of improved segments for the bilingual set-up for fidelity is, we suggest, that users were able to extract some of the meaning that was lost in the machine translation process from the source text. The fidelity scores for the French bilingual set-up did not increase much more than the fidelity scores for the French monolingual set-up. While there was a great improvement compared to the raw MT output, the fact that the values are very similar for both the monolingual and the bilingual set-up may be due to the fact that the participants' level of English did not make a difference in extracting more meaning for the bilingual set-up.

4.3 Per user - Detail

Figure 3 gives an overview of the average time spent in seconds per German participant per word split by set-up (monolingual and bilingual). It can be seen that whether more time is spent on monolingual or bilingual tasks varies across the post-editors. This could relate to the English skills of the participants. For example, participant B spent considerably more time on bilingual tasks, which may be explained by their knowledge of English - "5" (cf. Table 4) and was thus working more with reference to the source text than others.

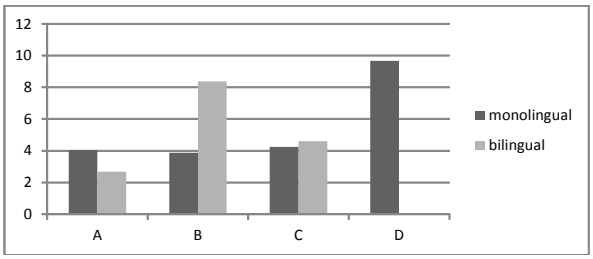

Figure 3: Time spent editing for each set-up (German) with time in average seconds per word

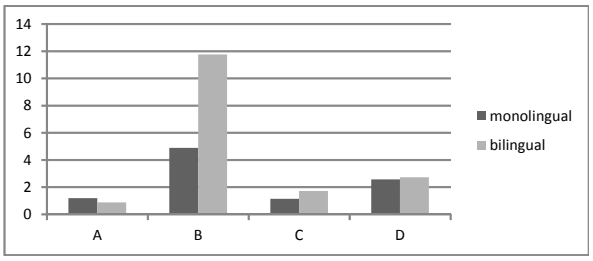

Figure 4: Time spent editing for each set-up (French) with time in average seconds per word

Figure 4 shows the time (on average in seconds per word) spent editing for each set-up for the French participants. Initially, it is striking that the French participants spent a considerably lower amount of time per word than the German participants, apart from one participant, who produced the most improved segments for all categories. This suggests that time may have a positive effect on all categories. The scarcity of data from this pilot study does not allow for a clear interpretation of the impact of time on post-editing quality, nor does the impact of task type on time.

4.4 Observations

While the evaluation strategy presented above gives a general overview of what impact post-editing can have on quality, an in-depth manual analysis of post-edited segments can give further insights into the issues encountered (cf., for example, Koponen 2011).

The first example displayed in Table 8 shows how a lack of fluency can be improved if com-

	ST	MT + back translation	PE + back translation
1.	if so how do I do **that**.	Wenn ja, wie soll ich tun. *If yes, how should I do.*	Wenn ja, wie soll ich **das** machen? Wenn ja, wie soll ich **das** tun. Wenn ja - wie soll ich **es** tun. *If yes, how should I do **that**?* Wenn ja, wie soll ich vorgehen? *If yes, how should I proceed?*
2.	Thats what I was after.	Das, was ich nach. *This, what I after*	So dachte ich zumindest. *That's what I thought at least.*
3.	after doind that I had to restore my Windows 7 Professional x64 SP1 pc, as all the computer magazine websites I use daily begin only showing comments without the article's the comments were for.I used Windows 7's snipping Tool for the screenshot's The forum won't upload either screenshot, which are both .png files.	Nach doind wiederherstellen, musste ich meine Windows 7 Professional X 64 SP1 pc, wie alle Computer Zeitschrift Website verwenden, die ich tglich nur Bemerkungen beginnen, ohne die Artikel der Kommentare waren for.I verwendet Windows 7 der Abstriche gemacht htte Tool fr die Screenshot Der Forum nicht entweder Screenshot hochzuladen, die sowohl .png Datei.	(backtranslation of MT:) *After doind restore, I had to my Windows 7 Professional x64 SP1 pc, as all the computer magazine website use, which I daily only comments begin, without the article of the comments were for. I used Windows 7 which would have been a compromise tool for the screenshot The forum not either upload screenshot, which either .png file.*
4.	RELATED: Any ideas where I (or my wife) might have picked these fun things up?	RELATED: jeder Ideen, wo ich meine Ehefrau) (oder vielleicht haben diese Spa machen? *RELATED: everyone ideas, where I my wife) (or maybe have these have fun?*	Keine Idee wo meine Ehefrau diese Dinge her hat. *No idea where my wife got those things from.*

Table 8: Post-editing examples (DE)

prehensibility and fidelity are high (2, 4, 4)[10]. All post-editors were able to compensate for the missing word "that". However, it is evident in this example that although all users were able to fix the error, they all opted for different solutions. The second example (scored 1, 0, 1) is an idiom. As the translation is incomplete and does not include a verb, it fails to communicate any meaning to a German speaker. This resulted in three participants leaving the segment as it is and one participant interpreting it freely, thus increasing fluency and comprehensibility scores but not improving fidelity (4, 4, 0). Such expressions do not deliver vital content in the forum posts but are essential to the individual writing style of the community members. For example three (0, 0, 0), one participant did not try to edit this, while three attempted to edit it. Two of them still scored 0 for fluency, comprehensibility and fidelity, while one participant deleted the content that was not understood and interpreted it based on the MT output, which resulted in a score of 3 for fluency, 4 for comprehensibility and 1 for fidelity. This is a very typical example of when post-editing is impossible, i.e. the information lost through MT cannot be retrieved from the machine translated text or compensated for by domain knowledge or other skills the users might have. In contrast to the second example, this segment is part of the problem description and is thus vital to the understanding of the user's problem. The poor MT output is here based on a poor source text including spelling mistakes, poor punctuation and complex sentences. It should be noted, however, that the availability of the ST does not automatically result in better results. For example four, one participant did not understand the ST fully, and while fidelity improved, this improvement was considerably below the fidelity scores of the other participants.

The misplacement of verbs (as in example two) and thus a loss of relation between the subject and the verb occurs quite frequently in the machine translation output of the current data and is a source of post-editing problems. Based on the data of the pilot study, the segments scoring low for both the MT output and the monolingually post-edited content can be traced back to mistakes in the ST, or colloquial or metaphorical language, which is something that may be addressed in a preprocessing step.

[10]These values indicate fluency, comprehensibility and fidelity scores (human evaluation)

5 Conclusion & Future Work

This study made a first attempt at uncovering whether forum users are able to improve raw MT output and whether the number of improved segments is greater than the number of degradations produced in a monolingual or bilingual post-editing environment. We found that there was a great variation between the post-editors' performance, especially for the German participants. It was evident that monolingual post-editing is not an unrealistic exercise, assuming forum users, for example, are willing to engage in it. When comparing the evaluated segments of the post-edited results with the evaluated segments of the raw MT output, we recorded a considerable increase in quality. What remains to be seen, however, is how factors such as language skills, domain knowledge (tested, rather than self-reported) and task time affect the quality in an experiment with a larger number of participants. For future studies, it would be desirable to include a larger number of participants, to make sure the participants understand the editing interface better to avoid loss of post-editing data, due to incorrect usage. With regards to the texts selected, the researchers were aiming at selecting similar texts that could be compared across the two set-ups (monolingual and bilingual). Unfortunately, direct comparability cannot always be guaranteed. Thus, an experiment with participants editing the same texts in different set-ups would allow for a more accurate comparison - but would require more participants. It would also be desirable to identify and investigate frequent changes made during the post-editing process in order to try to improve the SMT system. Furthermore, it would be preferable to include a larger number of human evaluators in order to obtain richer and more solid results.

Acknowledgements

This work is supported by the European Commission's Seventh Framework Programme (Grant 288769). The authors would like to thank Dr. Pratyush Banerjee for contributing the building of the clusters to group similar posts together for this post-editing study.

References

Callison-Burch, Chris and Koehn, Philipp and Monz, Christof and Schroeder, Josh, 2009. Findings of the 2009 Workshop on Statistical Machine Translation, *Proceedings of the Fourth Workshop on Statistical Machine Translation* 2009, Athens, Greece.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation* 2011, Edinburgh, U.K.

Giselle deAlmeida and Sharon O'Brien. 2010. Analysing Post-Editing Performance: Correlations with Years of Translation Experience. *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*

Marcello Federico, Nicola Bertoldi and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models *Interspeech 2008: 9th Annual Conference of the International Speech Communication Association*

Ignatius Garcia. 2010. Does Google know better? Translators and machine translation. *Translating and the Computer*, 32. 18-19 November 2010, London.

Anna Guerberof. 2009. Productivity and quality in MT post-editing, *MT Summit XII - Workshop: Beyond Translation Memories: New Tools for Translators, MT*, August 29, 2009, Ottowa, Ontario, Canada.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. 2009. The WEKA Data Mining Software: An Update, *SIGKDD Explorations*, Volume 11, Issue 1.

LDC. 2002. *Linguistic Data Annotation Specicication: Assessment of Fluency and Adequacy in Chinese English Translations*. Technical Report 1.0, Linguistic Data Consortium.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling, Volume 1 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

Philip Koehn. 2010. Enabling Monolingual Translators: Post-Editing vs. Options *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* Los Angeles, California: ACL.

Maarit Koponen. 2011. Correctness of machine translation: a machine translation post-editing task. *3rd MOLTO Project Meeting*, Helsinki.

Victor Muntes and Pat Paladini. 2012. Crowd Localization: bringing the crowd in the postediting process *Presentation at Translation in the 21st Century Eight Things to Change* Paris, May 31 - June 1 2012.

Donghui Lin, Yoshiaki Murakami, Toru Ishida, Yohei Murakami and Masahiro Tanaka. 2010. Composing Human and Machine Translation Services: Language Grid for Improving Localization Processes *Proceedings of Language Resources and Evaluation*, Valetta, Malta.

Johann Roturier and Anthony Bensadoun 2011. Evaluation of MT Systems to Translate User Generated Content *Proceedings of the 13th Machine Translation Summit (pp. 244251). Xiamen, China.*

Johann Roturier, Linda Mitchell and David Silva 2013. The ACCEPT Post-Editing environment: a flexible and customisable online tool to perform and analyse machine translation post-editing. *MT Summit XIV Workshop on Post-editing Technology and Practice. Nice, France.*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla and John Makhoul 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas*.

Midori Tatsumi and Takako Aikaw and Kentaro Yamamoto and Hitoshi Isahara 2012. How Good Is Crowd Post-Editing? Its Potential and Limitations. *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*.