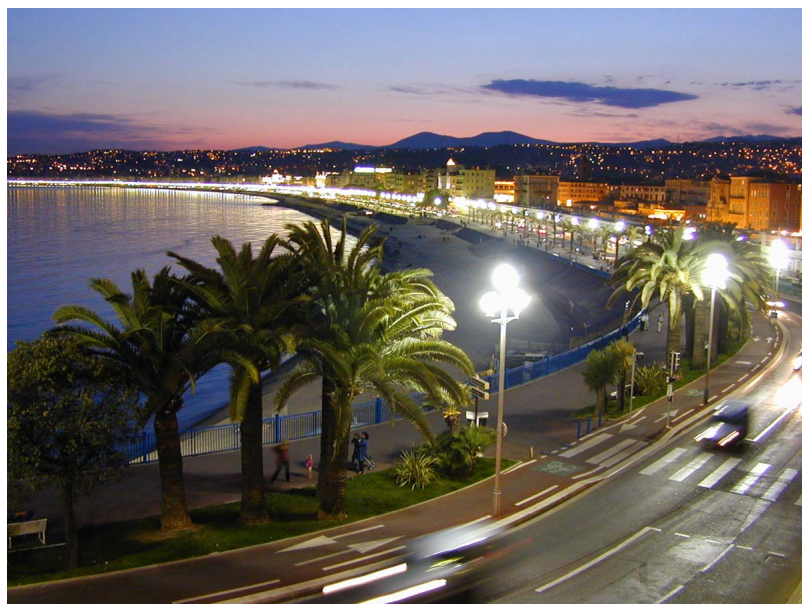


Machine Translation Summit XIV

2–6 September 2013, Nice, France



Workshop Proceedings: Workshop on Post-editing Technology and Practice (WPTP-2)

Editors: Sharon O'Brien, Michel Simard and Lucia Specia



Workshop Proceedings for:
The Workshop on Post-editing Technology and Practice
(Organised at the 14th Machine Translation Summit)

Editors: Sharon O'Brien CNGL/DCU
Michel Simard National Research Council Canada
Lucia Specia University of Sheffield

Published by
The European Association for Machine Translation
Schützenweg 57
CH-4123 Allschwil / Switzerland

ISBN: 978-3-9524207-2-0

©2013 The authors.

These proceedings are licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND

(For certain papers of these proceedings there may be stated other copyrights)

MT Summit 2013 Workshop Chair

Svetlana SHEREMETYEVA

Workshop Organisers

Sharon O'BRIEN – CNGL / Dublin City University
Michel SIMARD – National Research Council Canada
Lucia SPECIA – University of Sheffield

Program Committee

Nora ARANBERRI – TAUS
Diego BARTOLOME – tauyou language technology
Michael CARL – Copenhagen Business School
Francisco CASACUBERTA – Universitat Politècnica de València
Mike DILLINGER – eBay
Stephen DOHERTY – Dublin City University
Andreas EISELE – European Commission
Jakob ELMING – Copenhagen Business School
Marcello FEDERICO – FBK Trento, Italy
Mikel L. FORCADA – Universitat d'Alacant
Ana GUERBEROF – Pactera Technology International Ltd
Nizar HABASH – Columbia University
Kristian Tangsgaard HVELPLUND – Copenhagen Business School
Maxim KHALILOV – TAUS
Philipp KOEHN – University of Edinburgh
Roland KUHN – National Research Council Canada
Philippe LANGLAIS – RALI / Université de Montréal
Alon LAVIE – Carnegie Mellon University
Elliott MACKLOVITCH – MT Consultant
Daniel MARCU – SDL and USC/ISI
John MORAN – Transpiral Translation Services
Kristen PARTON – Facebook
Maja POPOVIĆ – DFKI
Johann ROTURIER – Symantec
Midori TATSUMI – Dublin City University
Jörg TIEDEMANN – Department of Linguistics and Philology, Uppsala University
Andy WAY – Lingo24

Workshop Programme

9:00 – 10:30: Session 1

9:00 Word of welcome

9:15 *This translation is not too bad: An analysis of post-editor choices in a machine translation post-editing task*
Maarit Koponen

9:40 *What can we learn about selection mechanism for post-editing?*
Maja Popović, Aljoscha Burchardt, Eleftherios Avramidis and David Vilar

10:05 *User Attitudes to the Post-Editing Interface*
Joss Moorkens and Sharon O'Brien

Coffee Break: 10:30 – 10:45

10:45 – 12:00: Session 2

10:45 *Integrated Post-Editing and Translation Management for Lay User Communities*
Adrian Laurenzi, Megumu Brownstein, Anne M. Turner and Katrin Kirchhoff

11:10 *Community-based post-editing of machine-translated content: monolingual vs. bilingual*
Linda Mitchell, Johann Roturier and Sharon O'Brien

11:35 *Combining pre-editing and post-editing to improve SMT of user-generated content*
Johanna Gerlach, Victoria Porro, Pierrette Bouillon and Sabine Lehmann

12:00 – 13:30: Lunch Break

13:30 – 14:30: Invited Talk

Organizational Ergonomics of Post-Editing
Mirko Plitt

14:30 – 16:25: Posters and Demos

Advanced Computer Aided Translation with a Web-Based Workbench
(Poster and demo)

Vicent Alabau, Ragnar Bonk, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes Garcia-Martinez, Philipp Koehn, Luis Leiva, Bartolome Mesa-Lao, Herve Saint-Amand, Chara Tsoukala, German Sanchis, Daniel Ortiz and Jesus Gonzalez

Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE (Poster)

Joke Daems and Lieve Macken

Transferring Source Tags to the Target Text in Statistical Machine Translation: A Two-Stream Approach (Poster)

Eric Joanis, Darlene Stewart, Samuel Larkin and Roland Kuhn

Assessing Post-Editing Efficiency in a Realistic Translation Environment (Poster)

Samuel Lübli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow and Martin Volk

Integrated Post-Editing and Translation Management for Lay User Communities (Demo)

Adrian Laurenzi, Megumu Brownstein, Anne M. Turner and Katrin Kirchhoff

VistaTec (Demo)

Phil Ritchie

The ACCEPT Post-Editing environment: a flexible and customisable on-line tool to perform and analyse machine translation post-editing (Demo)

Johann Roturier, Linda Mitchell and David Silva

PET: A tool for evaluating translation quality through post-editing (Demo)

Lucia Specia

An Evaluation of Tools for Post-Editing Research: The Current Picture and Further Needs (Poster)

Lucas Nunes Vieira

Online production of HQ parallel corpora and permanent task-based evaluation of multiple MT systems: both can be obtained through iMAGs with no added cost (Poster and demo)

Lingxiao Wang and Christian Boitet

Reverso (Demo)

16:25 – 17:30: Session 3

- 16:25 *Issues in Incremental Adaptation of Statistical MT from Human Post-edits*
Mauro Cettolo, Nicola Bertoldi, Marcello Federico,
Christophe Servan, Loïc Barrault and Holger Schwenk.
- 16:50 *The ACCEPT Post-Editing environment: a flexible and customisable on-line tool to perform and analyse machine translation post-editing*
Johann Roturier, Linda Mitchell and David Silva
- 17:15 Conclusion

Table of Contents

<i>This translation is not too bad: An analysis of post-editor choices in a machine translation post-editing task</i>	
Maarit Koponen	1
<i>What can we learn about selection mechanism for post-editing?</i>	
Maja Popovic, Aljoscha Burchardt, Eleftherios Avramidis and David Vilar	11
<i>User Attitudes to the Post-Editing Interface</i>	
Joss Moorkens and Sharon O'Brien	19
<i>Integrated Post-Editing and Translation Management for Lay User Communities</i>	
Adrian Laurenzi, Megumu Brownstein, Anne M. Turner and Katrin Kirchhoff	27
<i>Community-based post-editing of machine-translated content: monolingual vs. bilingual</i>	
Linda Mitchell, Johann Roturier and Sharon O'Brien	35
<i>Combining pre-editing and post-editing to improve SMT of user-generated content</i>	
Johanna Gerlach, Victoria Porro, Pierrette Bouillon and Sabine Lehmann	45
<i>Advanced Computer Aided Translation with a Web-Based Workbench</i>	
Vicent Alabau, Ragnar Bonk, Christian Buck, Michael Carl, Fran- cisco Casacuberta, Mercedes Garcia-Martinez, Philipp Koehn, Luis Leiva, Bartolome Mesa-Lao, Herve Saint-Amand, Chara Tsoukala, German Sanchis, Daniel Ortiz and Jesus Gonzalez	55
<i>Quality as the sum of its parts: A two-step approach for the identification of trans- lation problems and translation quality assessment for HT and MT+PE</i>	
Joke Daems and Lieve Macken	63
<i>Transferring Source Tags to the Target Text in Statistical Machine Translation: A Two-Stream Approach</i>	
Eric Joanis, Darlene Stewart, Samuel Larkin and Roland Kuhn	73
<i>Assessing Post-Editing Efficiency in a Realistic Translation Environment</i>	
Samuel Läubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow and Martin Volk	83
<i>An Evaluation of Tools for Post-Editing Research: The Current Picture and Fur- ther Needs</i>	

Lucas Nunes Vieira	93
<i>Online production of HQ parallel corpora and permanent task-based evaluation of multiple MT systems: both can be obtained through iMAGs with no added cost</i>	
Lingxiao Wang and Christian Boitet	103
<i>Issues in Incremental Adaptation of Statistical MT from Human Post-edits</i>	
Mauro Cettolo, Nicola Bertoldi, Marcello Federico, Christophe Servan, Loïc Barrault and Holger Schwenk.	111
<i>The ACCEPT Post-Editing environment: a flexible and customisable online tool to perform and analyse machine translation post-editing</i>	
Johann Roturier, Linda Mitchell and David Silva	119

Welcome from MT Summit 2013 Workshop Chair

On behalf of the hosts of the 14th Machine Translation Summit I am delighted to welcome our participants to Nice for the 2nd Workshop on Post-Editing Technologies and Practice.

In recent years we have witnessed a significant increase in popularity of machine translation post-editing. It has a great potential in developing methodologies and computer tools for correcting the raw MT output to improve the quality of translation. MT post-editing is inherent to MT as such and it is very important that a special full-day workshop devoted to this issue runs in conjunction with the current MT Summit to further promote up-to-date achievements in this field.

I should like to take this opportunity to thank Sharon O'Brien, Michel Simard and Lucia Specia, the organisers of the Workshop, whose enthusiasm and commitment made this event a high standard scientific gathering with a wide range of activities like oral presentations, discussions, poster and demo sessions.

I am sure you will find the Workshop a stimulating opportunity for sharing knowledge and expertise, meeting colleagues and friends.

Enjoy the Workshop and the gorgeous city of Nice!

Svetlana Sheremetyeva.

This translation is not too bad: An analysis of post-editor choices in a machine translation post-editing task

Maarit Koponen

University of Helsinki, Dept of Modern Languages

P.O. Box 24

00014 Helsingin yliopisto

`maarit.koponen@helsinki.fi`

Abstract

Variation between post-editors of machine translation is a well-known issue. This variation shows itself in post-editing speed, amount of editing and differing final translations. However, relatively few studies exploring the differences have been reported. This paper describes a post-editing task involving controlled language tourist phrases translated from English into Finnish. Post-editors select the best out of three machine translated suggestions, which they can accept without editing or post-edit as necessary. Agreement between editors is analyzed and reported in terms of selecting the best suggestion, deciding its acceptability, and producing a final post-edited version. Editors are compared in terms of post-editing time, edit distance and final translations created. With a qualitative analysis, we examine differences between the selected and rejected suggestions as well as differences between the post-edited versions created by different editors. Examples of editor preferences are also discussed.

1 Introduction

The growing interest in, and use of, machine translation (MT) post-editing as a way to increase productivity in professional translation scenarios has also recently led to growing interest on the research side. Tools and practices for post-editing (PE) are being developed, and PE tasks are being used to evaluate MT quality. A recognized issue in post-editing scenarios is the variation between

different editors, which shows in the amount of editing, PE speed and differing final translations produced. Post-editing, like translation in general, is an inherently subjective task in that the source meaning can generally be expressed in the target language in more than one way.

In analysing the variation between post-editors, attention has generally focused on questions of productivity: PE time and the technical effort of post-editing measured as keystrokes or edit distance between the MT and PE version (Krings, 2001; Plitt and Masselot, 2010; Tatsumi and Roturier, 2010; Koponen et al., 2012). Some studies have included analysis of the numbers of PE versions created and PE versions preferred by evaluators (Tatsumi et al., 2012), or examples of differing PE versions (O'Brien, 2005). However, much of the variation in post-editor choices and preferences as well as the factors influencing these choices still remains to be investigated.

In this paper, we aim to take some steps toward exploring the variation between editors in terms of the amount of editing performed (number of sentences edited and edit distance) and PE speed. We also examine the agreement between editors in selecting the best MT suggestion and deciding on whether to edit or not. For this purpose, we analyze data collected during a post-editing task involving a multilingual, controlled language generation and machine translation tool. The material, generated according to the controlled language rules, consists of short, relatively simple "tourist phrases" with limited vocabulary and structures: for example, questions about prices and directions, or small talk phrases. This type of material was selected for this study for its simplicity. The short, controlled sen-

tences were expected to lead to a relatively small number of MT errors, which would decrease the need for extensive rewriting and help to isolate the post-editing choices by different editors.

The post-editing task described in this paper involves the editors selecting the best out of three MT suggestions, accepting it without modification or post-editing as necessary. With this data, we set out to investigate the choices made by the editors in which suggestion to select and whether to accept it as such or edit. We will examine agreement between the editors and variation between different editors as well as the different PE versions. Using qualitative analysis, we will examine some of the preferences shown by the editors.

This paper is organized as follows. Section 2 presents prior research related variation in MT suggestion selections and post-editing. Section 3 describes the material and methods used in the analysis. Section 4 presents the analysis results. Conclusions from this study as well as future work are discussed in Section 5.

2 Related work

Selection of the best MT suggestion has often been used in large MT evaluation campaigns, such as those organized in context of the annual Workshop on Statistical Machine Translation (WMT) (Callison-Burch et al., 2012), where evaluation of competing systems involved the ranking of alternate MT suggestions. The results mostly focus on evaluating the quality of different systems, and do not generally include analysis of the differences between high and low ranking suggestions.

Prior MT post-editing studies have included analyses of the differences in PE speed and edit distances between different post-editors. An extensive analysis of increased productivity in post-editing compared to translation was carried out by Plitt and Masselot (2010) in a study involving twelve professional translators and various language pairs. Significant variation in post-editing speed was observed between post-editors.

In a study involving nine professional translators post-editing English-to-Japanese MT, Tatsumi and Roturier (2010) found that the translators differed more in terms of editing time than textual changes.

Koponen et al. (2012) report an analysis of human variability in post-editing. Eight post-editors editing English-to-Spanish MT were compared in

terms of PE time, keystrokes during editing, and edit distance. Post-editors were found to differ more in terms of PE time and keystrokes than edit distances.

Tatsumi et al. (2012) report the frequencies of multiple successive PE versions occurring in a scenario where post-edited versions of MT in various language pairs are crowdsourced from student participants. Most sentences are found to have no more than one translation version. Different crowdsourced versions are subsequently evaluated by professional translators who could either accept or revise them, and results are reported comparing whether the last or some earlier PE version is accepted.

One approach to studying different translation or post-editing choices is Choice Network Analysis (CNA). CNA has been suggested by Campbell (2000a,b) as a way to compare different translation versions created by multiple translators for a given source string. In Campbell (2000b), CNA is used to examine the behaviour of nine students related to two specific structures (cross-clause ellipsis and relative clauses).

O'Brien (2005) presents examples of using CNA to analyze the translations of four translator students post-editing English-to-German MT. Results obtained with CNA are compared to post-editing data, and long pauses in editing are found to correlate with locations indicated as difficult by CNA.

Blain et al. (2011) introduce the concept of Post-Edit Action (PEA), which combines multiple edit operations to linguistically logical groups, and analyze post-editing changes made by four professional translators on English-to-French MT. No comparison of individual editors' choices are reported.

The purpose of this study is to explore some aspects of agreement or disagreement between post-editors. Rather than post-editing times and edit distances, we focus on the agreement in selecting the best translation suggestion and in deciding acceptability. Further, we examine differences between individual editors and the final PE versions they create.

3 Material and analysis methods

The material analyzed for this paper consists of 139 sample sentences and their translations ob-

tained from the pilot evaluation material of a multilingual, controlled language text generation and machine translation system developed as part of the European MOLTO project¹. The dataset, described in more detail in Rautio and Koponen (2013), includes English source sentences, three Finnish MT versions of each sentence, and post-editing data from 11 post-editors. The total number of source words is 827, and source sentence length varies from 2 to 15 words per sentence (median 5).

The MT versions have been produced with the rule-based generation and translation tool in question, as well as the statistical MT systems Google Translate² and Bing Translator³. In some cases, multiple systems had produced the same MT suggestion. All systems produced an identical suggestion for 8 sentences, and two systems produced identical suggestions for 41 sentences. For 90 out of 139 sample sentences, three different MT suggestions were provided.

The post-editing data was collected using the open-source online MT evaluation tool Appraise (Federmann, 2012). The sentences were post-edited by 11 translator students who were native speakers of Finnish. Each editor was shown the 139 English source sentences together with the three Finnish MT suggestions. The order of sentences and suggestions was randomized by the evaluation tool. The editors were asked to select the MT suggestion they considered best and accept it as-is or post-edit as necessary. They were instructed to make only minimal corrections necessary. The option to create a translation from scratch was also given. In total, the dataset contains 1527 final translations created by the editors either by accepting or editing the MT suggestions. No sentences had been translated from scratch.

As the original sample sentences had been generated for the testing the rule-based system that was used to produce one of the MT versions, suggestions by this system can be expected to have an advantage in the selection. However, for the purposes of this study, we are interested in cases of agreement or disagreement between editors rather than the relative success of the systems.

The analysis of agreement between the editors

involves two issues: whether they agree on the selection of the best MT suggestion, and whether they accept the suggestion as-is or edit it. Combining these aspects leads to the following six possible scenarios:

1. The same MT suggestion is selected by all.
All accept without editing.
2. The same MT suggestion is selected by all.
None accept without editing.
3. The same MT suggestion is selected by all.
Some accept without editing.
4. Different MT suggestions are selected.
All accept without editing.
5. Different MT suggestions are selected.
None accept without editing.
6. Different MT suggestions are selected.
Some accept without editing.

Using the collected post-editing data, all sentences were categorized according to these scenarios. The final PE versions created by the editors were compared to calculate the number of different versions created for each source sentence. The most common PE version was also recorded.

Differences between individual editors were examined in terms of the amount of editing, PE time and agreement with the most common choices across all editors. For comparing the amount of editing, the Human-targeted Translation Edit Rate (HTER) was calculated using TERplus (Snover et al., 2009). The HTER score is calculated as the number of edit operations (word insertions, deletions, substitutions or word order shifts) between the MT and PE version divided by the number of words in the PE version. A HTER score of 0 indicates no changes while 1 indicates complete rewriting.

Sentence-level PE time automatically recorded by the evaluation tool was used for time comparisons. Information about the editors' choice of MT suggestion was compared to the most common selection for each sentence. Similarly, the editors' final version was compared to the most common version for each sentence.

Finally, the MT suggestions and final versions were analyzed manually by a native Finnish speaker. To investigate why some MT suggestions

¹<http://www.molto-project.eu/>

²<http://translate.google.com>

³<http://www.bing.com/translator>

were preferred over others, the selected and rejected MT suggestions were assessed for the correctness of meaning and language on a strict binary scale (fully correct/not fully correct). Cases where multiple PE versions had been created were compared to examine the differences between these versions.

4 Analysis results

This section presents the analysis results. Overall agreement between editors in terms of the MT suggestion selected and choice to accept or edit is presented in Section 4.1. The comparison of individual editors is presented in Section 4.2. The qualitative analysis of differences in MT suggestions and PE versions are discussed in Section 4.3.

4.1 Agreement between editors

	MT suggestions selected		
	Same	Different	Total
All accept	44	15	59
None accept	1	5	6
Some accept	33	41	74
Total	78	61	139

Table 1: Number of sentences categorized according to editor selections of same or different MT suggestions and choice to accept or edit.

Table 1 shows the distribution of cases into the six selection/acceptance categories. The columns show the number of sentences categorized by whether all editors selected the same MT suggestion or whether different suggestions were selected, as well as the total. The rows show the number of sentences categorized by whether the suggestion chosen by each editor was accepted by all, none or some editors.

Overall, the editors appear to mostly agree on which suggestion they select. When the 8 cases with three identical MT suggestions are excluded, all 11 editors select the same MT suggestion for 70 out of the remaining 131 source sentences (53%). In a further 29 cases (22%), only one editor selects a different option. This leaves 32 sentences (24%) where two or more people select a different suggestion. Only one case was found where each of the three MT suggestions were selected as best by at least one editor.

When the editors agree on the same MT suggestion, it is most often (44 sentences, 56.4%) accepted without editing. The cases where all editors found some suggestion acceptable, but disagreed on which one, were less common. For these 15 sentences, it appears that two of the suggestions are correct although different. Overall, there were only 6 cases where none of the editors found any MT suggestion acceptable. In one case they agree on which is the easiest to correct, whereas in the other five cases, different MT suggestions are selected.

The remaining cases represent a mixed situation where some accept and some edit. When one MT suggestion is selected by all (33 sentences), this suggestion still appears to be superior, but the editors disagree on whether it can be accepted as such. On the other hand, when the editors select different MT suggestions with some accepting and others editing (41 sentences, 67.2% of the cases where selections of best suggestion are split), there seems to be even more disagreement: some are willing to accept one suggestion while others rather edit a different one. Some potential reasons for these preferences are discussed in Section 4.3.

Table 2 shows the numbers of different PE versions produced for a given source sentence (1, 2, 3 or more than 3 versions). In addition to the total number of sentences, the rows show the number of sentences divided into the six defined selection/acceptance scenarios.

		Number of PE versions				
		1	2	3	≥ 4	Total
Same MT	All accept	44	0	0	0	44
	None accept	0	0	0	1	1
	Some accept	0	21	7	5	33
Different MT	All accept	0	15	0	0	15
	None accept	1	0	3	1	5
	Some accept	3	11	13	14	41
Total		48	47	23	21	139

Table 2: Number of different PE versions created by editors in the six selection/agreement scenarios.

Overall, most of the 139 sentences have only one or two final PE versions. For 48 sentences, only one final version was found. Nearly all of them (44 sentences) naturally relate to the cases where all editors have accepted the same version without modification. In one instance, all editors ended up with the same PE version although one of them started with a different MT suggestion, and

in three cases, one editor chose to edit a different suggestion but produced a PE version identical to the MT suggestion that was accepted by the other editors.

Cases with two different final versions were mostly produced by some editors accepting and others editing the same MT suggestion (21 sentences) or different editors accepting different suggestions (15 sentences). The remaining 11 cases involve situations where different suggestions are selected with varying acceptance or editing. For sentences with more than two versions, most also result from different suggestions being selected and varying choices whether to accept or edit. The highest number of different PE versions found was 10 (1 sentence).

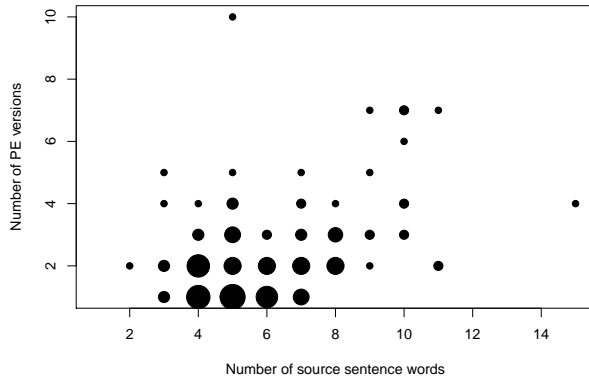


Figure 1: Plot showing the number of PE versions plotted against the number of source words. The circle size indicates multiple sentences with same values.

Figure 1 shows the number of PE versions for each sentence plotted against the number of source sentence words, with larger circles representing multiple sentences with the same value. All 48 cases with only one PE version, and nearly all with two versions (37 out of 47), involve sentences with 7 words or less. Most longer sentences, on the other hand, have 3 or more PE versions. With the exception of the one 5-word sentence with 10 different versions, sentences with the highest number of PE versions have more than 8 words. This is likely at least partly due to the shorter sentences having better MT quality and more often being accepted as-is. Conversely, longer sentences contain more errors and more need for editing then leads to more variation in the solutions found by different editors.

4.2 Comparison of individual editors

Figure 2 shows the number of sentences edited by each editor. Overall, all editors mostly accept one of the suggestions as-is. The number of sentences edited by each editor ranges from 19 (13.7% of all sentences, FI11) to 46 (33.1%, FI01) with a median of 39 sentences (28.1%).

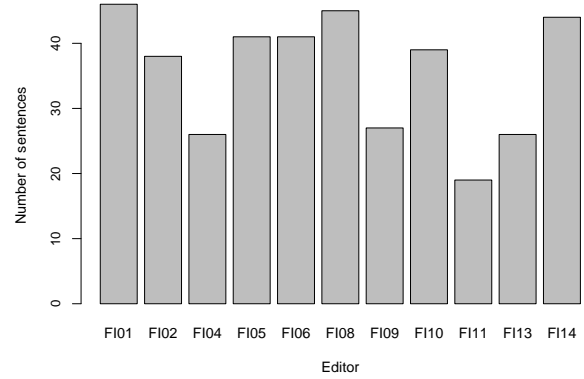


Figure 2: Bar plot showing the number of sentences edited by each editor.

Figures 3(a) and (b) show box plots of edit distances by editor. In the box plots, the bottom and top of the box represent the first and third quartile, and the line inside the box shows the median. The whiskers extend to 1.5 times the length of the box, and individual circles represent the cases outside of these limits.

Figure 3(a) shows the edit distance for all sentences. Because all editors accepted the majority of sentences without editing (see Figure 2), the median HTER score for each editor 0 is in Figure 3(a). To provide a clearer picture of how much each editor edited when they *did* decide editing was necessary, Figure 3(b) shows the edit distances for only those sentences that had been edited. The low HTER scores indicate that even when editing is considered necessary, a relatively small number of changes is made. Overall, there do not appear to be great differences between the editors, as all editors have median HTER between 0.17 and 0.20 except FI04 (median 0.23).

Figure 4 shows box plots of PE times by editor. One outlier sentence (from editor FI13) with a PE time over 300 seconds was removed, which was the only case where PE time exceeded 100 seconds. Median times are between 5.4 and 10.0 seconds per sentence for all editors except

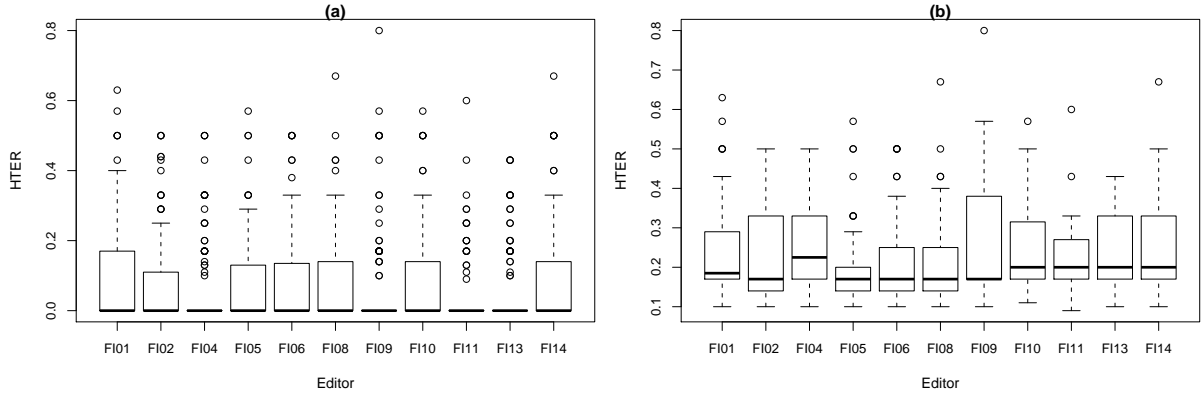


Figure 3: Box plots showing the edit distances (HTER) for each editor. HTER scores are shown for all sentences (a) and for edited sentences only (b).

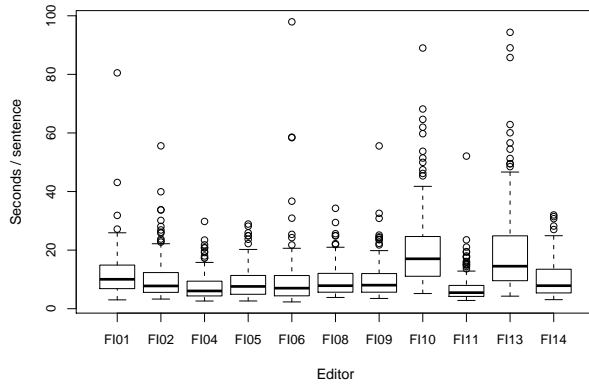


Figure 4: Box plots showing the editing times for each editor.

FI13 (14.5 seconds/sentence) and FI10 (17.0 seconds/sentence). For the three fastest editors (FI04, FI05, FI11), even the slowest times are around 30 seconds per sentence.

The editors were also compared in terms of how often their choice of best MT suggestion differed from the majority and how often they produced a final PE version differing from the most common version.

Figure 5 shows a bar plot of the number of sentences where each editor selected a different MT suggestion than the majority. The number of cases where each disagreed with the majority varies between 6 (FI01) and 17 (FI13), with median of 10.

Figure 6 shows a bar plot of the number of cases where each editor produced a PE version different from the majority. The number of such differing versions ranges from 19 (FI09) to 35 (FI02), with a median of 29.

Comparing these figures, some editors appear

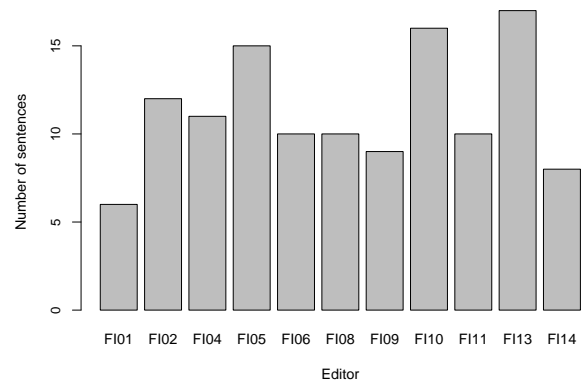


Figure 5: Bar plot showing the number of sentences where each editors' selection of MT suggestion differs from the majority.

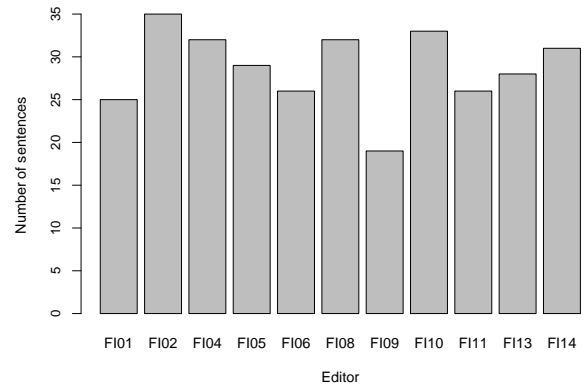


Figure 6: Bar plot showing the number of sentences where each editors' PE version differs from the majority.

to stand out: Editor FI04 has one of the smallest numbers of edited sentences and is also one of the fastest, but seems to edit slightly more than the oth-

ers and is among those who most commonly produce a PE version differing from the majority. On the other hand, editor FI11 edits relatively few sentences, but is one of the slowest and rarely deviates from the most common PE version. Editors FI01 and FI09 appear to commonly agree with the majority, as both have low numbers of selections and PE versions differing from the majority. These observations of differing profiles share some similarities with results reported in Koponen et al. (2012).

4.3 Qualitative analysis of editor preferences

Section 4.1 presented results of how often the editors agreed in selecting the best MT suggestion and whether it needed further post-editing. In this Section, we aim to examine some possible explanations for their choices and differences by comparing, on one hand, the selected and rejected MT suggestions, and on the other hand, the differing PE versions.

In the cases where at least two MT suggestions were offered and all editors selected the same one (70 sentences), most rejected suggestions had neither correct meaning nor correct language. They generally contained multiple errors which sometimes made it difficult to ascribe any meaning to the sentence – for example, *Jossa on suosituin Puolan ravintolassa?* ‘In which is the most popular of Poland in the restaurant?’ for *Where is the most popular Polish restaurant?* Interrogative sentences commonly contained multiple MT errors involving missing interrogative suffixes and wrong word order, literal translations of *do* or incorrectly added negation, affecting both language and meaning.

There were 15 cases where the rejected MT suggestion had correct language but different meaning than the source. This occurred, for example, in possessive structures, where the wrong case in the possessor noun can lead to suggestions like *Hänen vaimonsa on maitoa.* ‘His wife is milk’ instead of *Hänen vaimollansa on maitoa.* ‘His wife has milk.’ Other changed meanings involved incorrect words. In 3 cases, the rejected MT suggestion was assessed to have correct meaning despite incorrect language. These involved sentences with incorrect subject-verb agreement that is not standard in written Finnish but commonly used in spoken language.

In 3 cases, the meaning and language of the re-

jected suggestions was correct, but all the editors still preferred another suggestion. In these cases, the editors appear to have made the decision based on specific words or expressions, such as *Oletko kahdeksanvuotias?* for ‘Are you eight years old?’ rather than *Oletko kahdeksan vuotta?* Other sentences with similar expressions and varying editor choices were found.

In the 15 cases where all editors have accepted some MT suggestion without editing but disagree on which one, the selected suggestions generally differ from each other in ways that leave both the meaning and language correct. They mostly involved choice between synonymous words or expressions, such as *avoinna* vs *auki* for ‘open’. As Finnish word order is relatively free, word order was also a recurring difference. One case of differing punctuation was also found.

In 4 cases, one of the selected versions had correct language but was not, in fact, precisely correct in terms of meaning. These sentences involved cases where Finnish makes a distinction not present in English: the pronoun *they*, where Finnish uses different words for humans and non-humans, or second person forms, where Finnish distinguishes between informal singular, polite singular, and plural. During the post-editing task, the English sentences were presented with disambiguation information, but some editors appear to have ignored this. Two cases where MT suggestions with both incorrect meaning and incorrect language were accepted by at least one editor were also found.

Differences in preferences could also be observed in the cases where the editors disagreed whether the same MT suggestion needed editing or not and produced differing PE versions. Some recurring differences involved punctuation, specifically commas between main and subordinate clauses (required, but commonly omitted particularly in short sentences), alternate spellings such as *pizza* vs *pitsa* ‘pizza’ or alternate suffixes such as *dollareja* vs *dollareita* ‘dollar (plural partitive)’, as well as synonyms.

At least some cases where the editors disagree on which MT suggestion to select and whether it needs editing appear to be connected to particularly strong preferences for specific words or expressions. One such preference involved the choice of the verb *tahtoa* or *haluta* ‘want’. Most

editors appear to have at least some preference for *haluta*, since options containing that word were generally selected by all or nearly all editors if otherwise correct, and MT suggestions containing the alternative *tahtoa* were often edited to change this verb even when otherwise correct. Seven cases were identified where at least one editor even chose to edit MT suggestions that had both incorrect language and different or unclear meaning but contained a form of the preferred verb *haluta*, rather than accept or even edit a (correct) version containing *tahtoa*.

Some sentences or expressions seemed to generate a large number of different PE versions. As mentioned in Section 4.1, one sentence received a total of 10 different versions: *This apple is not too bad*. Without context, it can be understood either literally ("bad, but not too much so") or idiomatically ("quite good"). Other sentences with the "not too (adjective)" structure have been interpreted literally by the editors, but in this case, all but one chose the idiomatic interpretation and expressed it with varying wording. Other cases leading to multiple PE versions involved sentences like *Do you know how far the park is by bike?* for which a total of seven different versions were created.

Such cases where particularly many versions were created are similar to findings obtained using Choice Network Analysis. CNA assumes that multiple target versions of a given source string indicate parts that are difficult cognitively, as no single obvious solution is available to the translator or post-editor (Campbell, 2000b). A connection with pauses during post-editing was reported in O'Brien (2005), supporting this assumption. The sentences involving *not too bad* and *how far by* may indeed have caused difficulty. However, for some of the differences, such as the word choice for translating *open*, the variation may simply indicate varying preferences without any particular difficulty.

5 Conclusions and future work

The purpose of this study was to examine editors involved in an MT post-editing task, their editing choices and agreement between editors.

For most source sentences, all or all but one editor select the same MT suggestion and most sentences only have one or two PE versions. This is likely to be related to the nature of the controlled

language, high MT quality, and the large number of suggestions accepted without modification. Some differences were found between individual editors in terms of the number of sentences edited, and how often they deviated from the most common selection or most common PE version. Similar to prior studies, less variation was observed in edit distances than in PE times.

As expected, the editors tended to reject MT suggestions with multiple errors leading to both incorrect meaning and language. Variation in the selection of best MT suggestion and final PE versions appeared to mainly relate to choice of specific words or expressions or the use of punctuation. Cases where some editors chose to edit an incorrect sentence over a version accepted as correct by others were also identified. Examples of editor preferences related to these choices were discussed.

The sample is rather limited due to the controlled language. However, the repeated vocabulary and structures offer a chance to observe editor choices across similar cases. In future work, we are interested performing a more quantitative analysis of factors potentially influencing the editors' choices. Working with a less controlled text type would likely reveal more variation in the editor's choices, and would therefore be desirable. Professional translators might also produce results different from translator students. One question to study would be whether the version containing the preferred words (even in incorrect form) or word forms (even if not preferred words) would be preferred by editors. For this purpose, automatic tagging and lexical resources such as WordNets could be used. Specific editor preferences could also be explored in more detail.

Acknowledgements

This work has been supported by LANGNET Finnish Graduate School in Language Studies. The author wishes to thank the EU MOLTO project, particularly Evaluation Coordinator Jussi Rautio and Professor Lauri Carlson, for the use of the evaluation material and for discussions.

References

Blain, Frédéric, Jean Senellart, Holger Schwenk, Mirko Plitt and Johann Roturier 2011. Qualitative analysis

- of post-editing for high quality machine translation. In *MT Summit XIII: the Thirteenth Machine Translation Summit*, Xiamen, China. 164–171.
- Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montréal, Canada. 10–51.
- Campbell, Stuart. 2000. Choice Network Analysis in Translation Research. In M. Olohan (ed.) *Intercultural Faultlines: Research Models in Translations Studies: Textual and Cognitive Aspects*. St. Jerome, Manchester. 29–42
- Campbell, Stuart. 2000. Critical Structures in the Evaluation of Translations from Arabic into English as a Second Language. *The Translator*, 6: 37–58.
- Federmann, Christian. 2012. Appraise: An Open-Source Toolkit for Manual Evaluation of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 98: 25–35
- Koponen, Maarit, Wilker Aziz, Luciana Ramos and Lucia Specia. 2012. Post-editing time as a measure of cognitive effort. *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP 2012)*, San Diego, USA. 11–20.
- Krings, Hans P. 2001. *Repairing texts: Empirical investigations of machine translation post-editing process*. G. S. Koby (ed.). The Kent State University Press, Kent, OH.
- O’Brien, Sharon. 2005. Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability. *Machine Translation*, 19(1):37–58.
- Plitt, Mirko and Francois Masselot. 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, 93: 7–16.
- Rautio, Jussi and Maarit Koponen. 2013. *D9.2 MOLTO evaluation and assessment report*. Technical report, MOLTO project May 2013. <http://www.molto-project.eu/biblio/deliverable/d92-molto-evaluation-and-assessment-report>.
- Snover, Matthew, Nitin Madnani, Bonnie Dorr and Richard Schwartz. 2009. Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.
- Tatsumi, Midori and Johann Roturier. 2010. Source Text Characteristics and Technical and Temporal Post-Editing Effort: What is Their Relationship? In *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC 10)*, Denver, USA. 43–51.
- Tatsumi, Midori, Takako Aikawa, Kentaro Yamamoto and Hitoshi Isahara. 2012. How Good Is Crowd Post-Editing? Its Potential and Limitations. *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP 2012)*, San Diego, USA. 69–77.

What can we learn about the selection mechanism for post-editing?

Maja Popović, Eleftherios Avramidis, Aljoscha Burchardt,
David Vilar, Hans Uszkoreit
DFKI / Berlin, Germany
name.surname@dfki.de

Abstract

Post-editing is an increasingly common form of human-machine cooperation for translation. One possible support for the post-editing task is offering several machine outputs to a human translator from which then can choose the most suitable one. This paper investigates the selection process for such method to get a better insight into it so that it can be optimally automatised in future work. Experiments show that only about 70% of the selected sentences are the best ranked ones, and that selection mechanism is tightly related to edit distance. Furthermore, five types of performed edit operations are analysed: correcting word form, reordering, adding missing words, deleting extra words and correcting lexical choice.

1 Motivation and related work

Machine translation (MT) has improved considerably in recent years thus gaining recognition in the translation industry. However, machine translation outputs have not yet reached the same quality as human translations. Performing the post-editing has become a common practice for improving machine translation outputs. Therefore, more and more attention is paid to various aspects of post-editing, such as (Specia, 2011). Prediction of errors in rule-based system outputs has been investigated in (Valotkaitė and Asadullah, 2012) in order to facilitate the post-editing process. Analysis of edit operations has been carried out in (Koponen, 2012) in order to understand discrepancies between

edit distance and translation quality (i.e. predicted post-editing effort).

Our work explores the selection criteria applied by professional translators when several translation outputs of each source sentence are offered for post-editing. The scenario is similar to the one in (He et al., 2010), but our approach goes beyond, since they consider only two outputs (one produced by statistical machine translation system and other by translation memory), they do not examine ranking of these outputs, they have not tested their automatic method by professional translators, and they do not analyse edit distances and the performed edit operations. Our main questions are:

- Is the translation output which is best for post-editing also the best ranked one?
- Is the edit distance of the chosen output lower than edit distances of the other outputs?
- Are there some (less) preferred edit operations?

and to the best of our knowledge they have not been investigated yet.

2 Experimental setup

The translation outputs investigated in this work are produced by German-English, German-French and German-Spanish machine translation systems in both directions. The test sets consist of three domains: news texts taken from WMT tasks (Callison-Burch et al., 2010), technical documentation extracted from the freely available OpenOffice project (Tiedemann, 2009) and client data owned by project partners. The number of

	News	OpenOffice	Client	Total
de-en	1788	418	500	2706
de-es	514	414	548	1476
de-fr	912	412	382	1706
en-de	1744	414	0	2158
es-de	101	413	1028	1542
fr-de	1852	412	0	2264
Total	6911	2483	2458	11852

Table 1: Test sets for ranking task and selecting for post-edit task – number of source sentences per language pair and domain.

source sentences per language pair and domain can be seen in Table 4.

Four translation systems were used: a phrase-based statistical machine translation (SMT) system Moses (Koehn et al., 2007), a hierarchical SMT system Jane (Vilar et al., 2010), a commercial rule-based system Lucy (Alonso and Thurmaier, 2003), and another commercial rule-based system RBMT¹.

The translation outputs generated by the described systems were then given to professional translators in order to perform ranking and post-editing using the browser-based evaluation tool Appraise (Federmann, 2010).

Ranking and post-editing tasks were defined as follows:

Ranking: for each source sentence (11852 sentences in total), rank the outputs of four different MT systems according to *how well these preserve the meaning of the source sentence*. Ties were allowed.

Select and post-edit: for each source sentence (11852 sentences in total), select the translation output *which is easiest to post-edit* and perform the editing.

Post-edit all: for each source sentence in the selected subset (4070 sentences in total), post-edit all four produced translation outputs.

For both post-editing tasks, the translators were asked to perform only the minimal post-editing necessary to achieve acceptable translation quality. Post-editing all translation outputs is a more

¹The system’s name is not mentioned here by request of the vendor.

rank	1	2	3	4
Overall	71.7	19.1	6.5	2.7
News	70.0	20.4	7.2	2.3
OpenOffice	62.3	24.4	8.0	5.2
Client	84.1	10.4	3.6	1.7
de-en	69.4	20.1	7.0	3.5
de-es	80.4	15.0	3.8	0.8
de-fr	68.0	21.1	8.1	2.8
en-de	66.3	22.1	8.9	2.7
es-de	77.4	15.5	3.8	3.3
fr-de	67.4	21.1	7.8	3.6

Table 2: Percentage of sentences with a given rank selected as the best for post-editing.

complex and time-consuming task in comparison to post-editing only the selected outputs, therefore only a subset of source sentences was selected.

3 Results

3.1 Selection vs. ranking

The first question we want to answer is how the sentences chosen for post-editing were ranked in the ranking task. Table 2 shows the percentage of selected sentences for each of four ranks (1 being the best, 4 the worst). It can be seen that overall, only 70% of selected sentences were ranked as best. About 20% of selected sentences were ranked as second best, and 10% had one of the two lowest ranks. For the client data, the percentage of the first ranked selected sentences is higher (84%) as well as for the language pair German–Spanish in both translation directions, and for the technical documentation is lower (62%). The results for the rest of domains and language pairs show the same tendency as the overall results.

Table 3 shows an example of a third ranked translation selected for post-editing extracted from German-to-English client data: one word remained untranslated which degraded significantly the quality. On the other hand, the correction of this sentence is easy – it requires only one edit operation, namely replacing this (German) word with the correct (English) one. This shows that the post-editor’s expectations about the amount of editing necessary, which could be approximated by the edit distance, are taken into account when it comes to select the translation to be post-edited.

source	Dazu ist ein Schraubendreher erforderlich.
Rank	Translation output
1	For this purpose a screwdriver is necessary.
2	In addition a screwdriver is necessary.
3*	This requires a Schraubendreher.
4	This would require an Schraubendreher required.
edit(3)	This requires a screwdriver.

Table 3: Example of discrepancy between ranking and post-editing: the third ranked sentence is chosen for post-editing due to lower edit distance.

3.2 Edit distances

The previous results show that there is a difference between the selection mechanisms for ranking translation outputs based on meaning and for choosing the output most suitable for post-editing. The results also confirmed that the edit distance plays an important role for the post-editing selection, but the further question is how exactly. It would be good to know if only the total edit distance matters, or some types of edit operations are more or less preferred than the others.

In order to explore these aspects, automatic edit analysis was carried out using the Hjerson tool (Popović, 2011) using the post-edited translations as references. The following five types of edit operations were distinguished: correcting word form (morphology), correcting word order, adding missing word, deleting extra word and correcting lexical choice. The results are presented in the form of edit rates, i.e. the total number of edit operations normalised over the total number of words. The total edit distance was calculated as a sum of the five edit rates.

3.2.1 Selected vs. rest

The first step in edit distance analysis was to compare edit distances of the selected sentences with the edit distance of the remaining sentences which were not selected. The obtained edit rates together with the relative differences ($\text{editRate}(\text{rest}) - \text{editRate}(\text{sel}) / \text{editRate}(\text{rest})$) are presented in Table 4. The first two columns show the edit rates for selected sentences and for the rest, and the

	edit rates (%)		relative difference (%)
	selected	rest	
form	2.9	4.5	36.2
order	5.3	7.8	31.9
missing	3.6	6.7	45.8
extra	6.0	9.0	34.2
lexical	21.2	33.0	35.8
total	39.0	61.0	36.0

Table 4: Total edit distance and five distinct types of edits (%) for selected sentences and not selected sentences (first row) and their relative differences (%) (second row).

third column presents their relative differences. Overall, the relative difference between the edit distances of two sets is 36%, meaning that 36% less edit operations were performed in the selected sentences than in the rest of the sentences. The relative differences are similar for all edit operation types being between 30% and 36%, except the missing words with 45% – adding missing words does not seem to be preferred in general. Further analysis is necessary for drawing definite conclusions.

We carried out a further analysis in a somewhat different direction, namely compare the selected sentences which are not best ranked with their best ranked "opponents".

3.2.2 Selected vs. best ranked

Further analysis was constrained only on the best ranked sentences which were not selected for post-editing. The first step was to calculate total edit distances of these sentences and their selected counterparts, and the results are presented in Table 5. Overall, the edit distance for the selected sentences is lower than for the best ranked confirming that the edit distance is a very important factor for the selection mechanism. Separate edit distances for three distinct domains show the same tendencies. However, the results for separated translation directions showed that there are exceptions – some selected sentences have higher edit distance than their best ranked "opponents". Two examples from the German-to-English task are shown in Table 6 and further analysis of such sentences is shown in the next section.

The first example shows a preference to two lexical corrections over one reordering. In the sec-

edit distance (%)	selected	rank 1
Overall	37.1	48.9
News	38.5	48.0
OpenOffice	33.3	51.3
Client	39.5	44.7
de-en	30.8	38.9
de-es*	35.9	33.9
de-fr	57.8	67.8
en-de*	44.9	37.6
es-de	32.8	51.7
fr-de	42.4	44.5

Table 5: Total edit distances (%) for selected sentences and best ranked not selected sentences: values are normalised over the total number of words.

source	Inzwischen sei das träge fließende Gewässer vollkommen tot .
rank 1	Now are _{reord} the lazy river waters completely dead .
edit	Now the lazy river waters are completely dead .
selected (rank 3)	Meanwhile the sluggish _{lex} river _{lex} waters are completely dead .
edit	Meanwhile the sluggishly flowing waters are completely dead .
source	Probleme gibt es auch bei Kilometer 185 der Autobahn D1 in Richtung Prag .
rank 1	There are problems _{reord} also _{reord} at kilometer 185 of the motorway D1 in direction Prague .
edit	There are also problems at kilometer 185 of the motorway D1 in the direction of _{miss} Prague .
selected (rank 2)	There are problems _{reord} also _{reord} with _{lex} kilometer of _{reord} 185 _{reord} the motorway D1 toward Prague .
edit	There are also problems at kilometer 185 of the motorway D1 toward Prague .

Table 6: Examples of not best ranked selected sentences with larger edit distance than their best ranked counterparts.

ond example, two reorderings and one lexical corrections are performed in the selected sentence whereas in the best ranked one there are only one reordering and one omission, suggesting again that translators tend to avoid adding missing words.

3.2.3 Selected vs. best ranked with lower edit distance

Comparison between edit distances of not best ranked selected sentences and their best ranked "opponents" in the previous section generated a new question: why the translators sometimes choose sentences which are neither the best translations nor the closest translations? Further edit rate analysis was constrained only on those sentences, namely the selected sentences which are neither best ranked nor have the lowest edit distance and the five edit rates for those sentences are graphically presented in Figures 1 and 2.

Figure 1 presents the overall five edit rates together with the edit rates for each of three domains. As expected, all edit rates are higher for the selected sentences, and furthermore it can be noted that the differences are largest for reordering edit rates. Only for the technical (OpenOffice) all differences are very small.

The results for different translation directions are shown in Figure 2, and it can be seen that the differences between edit rates are rather language-dependent, although a larger reordering edit rate for selected sentences can be observed for all translation directions. On the other hand, word form (inflection) edit rate of selected sentences is significantly higher only for the English-to-German translation. A possible explanation is that the German inflections often cannot be generated properly when translating from morphologically poorer English, however correcting them does not pose a big problem for translators, especially in comparison to other edit operations. Another interesting observation is that there are neither significant nor conclusive differences between the effects of missing words – a thorough analysis of this edit operation should be carried out, however it seems that although adding missing words is generally not the preferred action for the translators, it does not influence significantly the selection of a low(er) ranked sentence.

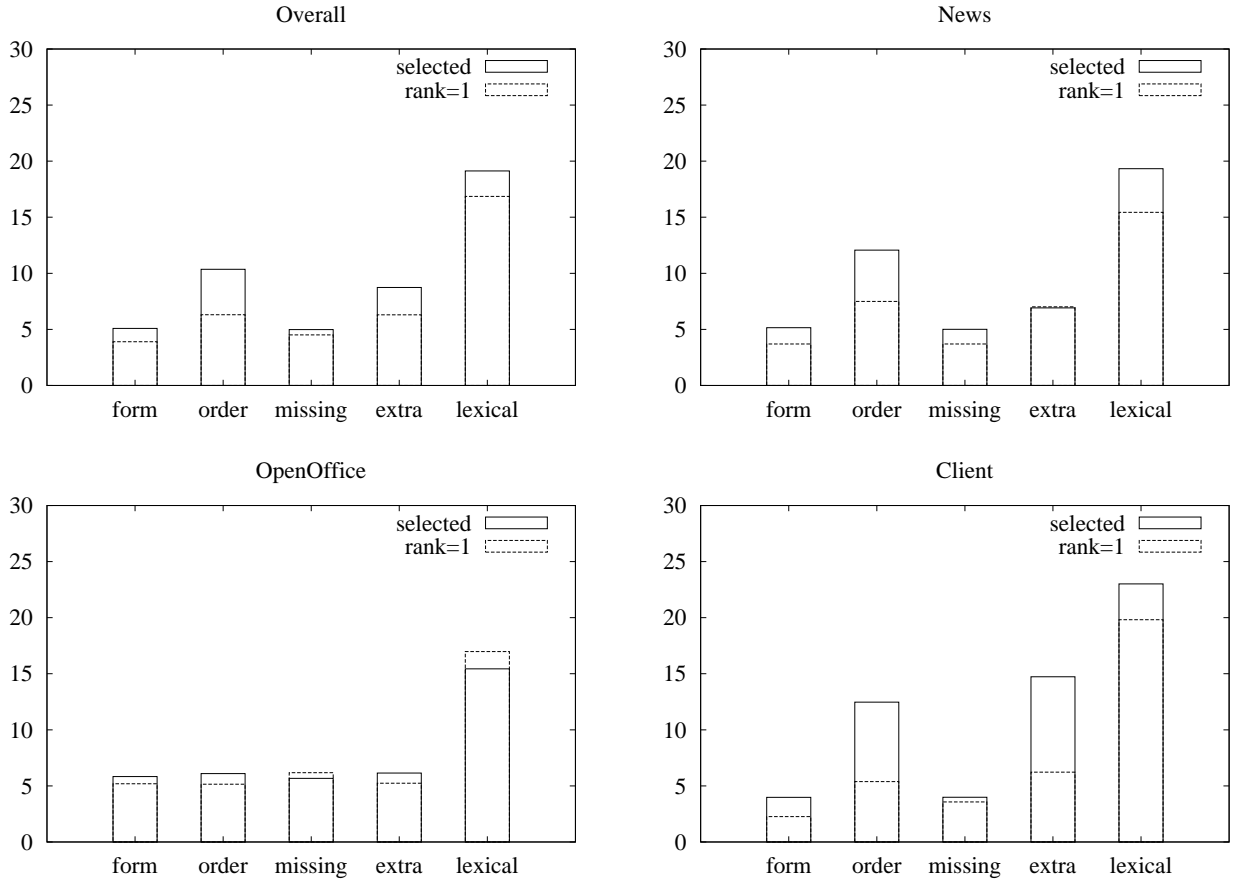


Figure 1: Five edit rates (%) of selected sentences with lower rank and larger edit distance (selected) and their best ranked counterparts (rank 1): overall and for three different domains separately.

Further analysis for different translation directions is carried out in the form of the distribution of edit operations over POS tags. For all language pairs, reordering of the noun is performed much more often in the selected sentences than in others. In addition, the amount of preposition reordering edit operations differs for all translations from German, whereas inflection and reordering of determiners are distinctive in all translations into German.

4 Summary and outlook

In this paper we investigated the post-editing selection mechanism of human translators by analysis of ranks, total edit distances and five types of edit operations. It is shown that only about 70% of the selected sentences are at the same time the best ranked ones, therefore the selection mechanisms for the best output and for the output best for post-editing differ significantly. Furthermore,

it is shown that the post-editing selection mechanism may be modelled in terms of the post-editor’s perception of the amount of post-editing needed, which may be measured a posteriori using the actual edit distance between the raw and the post-edited sentence. Nevertheless, a simple edit distance is not the only criterion. Further analysis has shown that some phenomena are rather language-dependent, however reordering edit operation is distinctive for all test sets. In addition, it is shown that reordering of nouns plays a significant role for all translation directions.

This work can be extended in various ways. One direction is using the obtained results for already mentioned automatisations of the selection process. Another direction is investigation of selection criteria for different translation systems, e.g. comparing statistical and rule-based systems. Furthermore, more detailed analysis including distinct types of edit operations and POS tags as well as

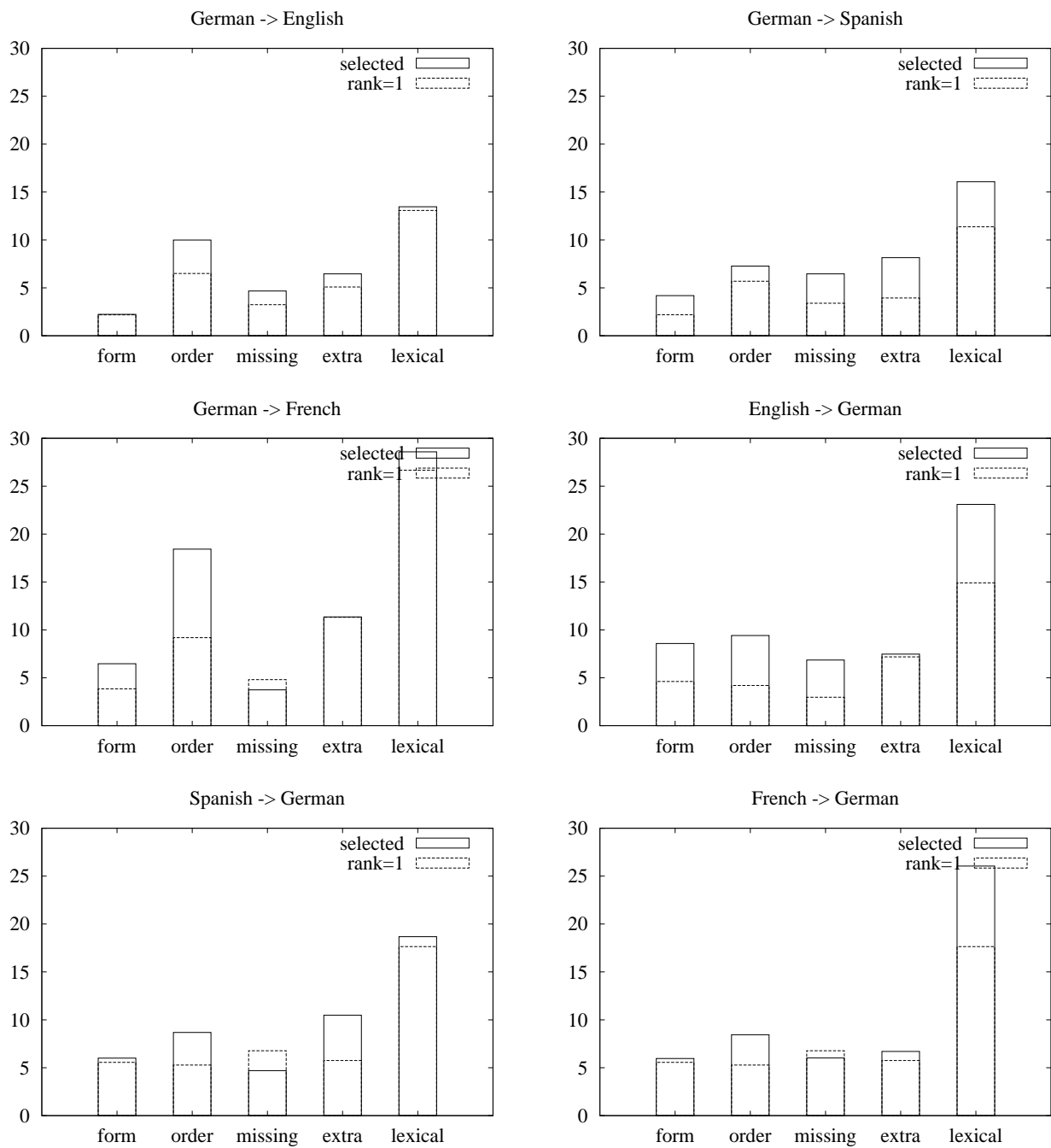


Figure 2: Five edit rates (%) of selected sentences with lower rank and larger edit distance (selected) and their best ranked counterparts (rank 1): the six translation directions are shown separately.

further investigation of missing words in various scenarios should be carried out on different language pairs and translation directions.

Acknowledgments

This work has been developed within the TARAXÚ project financed by TSB Technologiestiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development.

References

- Alonso, Juan A. and Gregor Thurmair. 2003. The comprehendium translator system. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, LA, September.
- Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT 2010)*, pages 17–53, Uppsala, Sweden, July.
- Federmann, Christian. 2010. Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, May.
- He, Yifan, Yanjun Ma, Josef van Genabith, and Andy Way. 2010. Bridging SMT and TM with translation recommendation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 622–630, Uppsala, Sweden, July.
- Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Chris Zens, Richard and Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07*, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Koponen, Maarit. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 181–190, Montréal, Canada, June. Association for Computational Linguistics.
- Popović, Maja. 2011. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, (96):59–68, October.
- Specia, Lucia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 11)*, pages 73–80, Leuven, Belgium, May.
- Tiedemann, Jorg. 2009. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Advances in Natural Language Processing*, volume V, chapter V, pages 237–248. Borovets, Bulgaria.
- Valotkaitė, Justina and Munshi Asadullah. 2012. Error Detection for Post-editing Rule-based Machine Translation. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, page 7886, San Diego, USA, October. Association for Machine Translation in the Americas (AMTA).
- Vilar, David, Daniel Stein, Matthias Huck, and Hermann Ney. 2010. Jane: Open source hierarchical translation, extended with reordering and lexicon models. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT 2010)*, pages 262–270, Uppsala, Sweden, July.

User Attitudes to the Post-Editing Interface

Joss Moorkens

CNGL/SALIS

Dublin City University

Ireland

joss.moorkens@dcu.ie

Sharon O'Brien

CNGL/SALIS

Dublin City University

Ireland

sharon.obrien@dcu.ie

Abstract

At present, the task of post-editing Machine Translation (MT) is commonly carried out via Translation Memory (TM) tools that, while well-suited for editing of TM matches, do not fully support MT post-editing. This paper describes the results of a survey of professional translators and post-editors, in which they chose features and functions that they would like to see in translation and post-editing User Interfaces (UIs). 181 participants provided details of their translation and post-editing experience, along with their current working methods. The survey results suggest that some of the desired features pertain to supporting the translation task in general, even before post-editing is considered. Simplicity and customizability were emphasized as important features. There was cautious support for the idea of a UI that made use of post-edits to improve the MT output. This research is intended as a first step towards creating specifications for a UI that better supports the task of post-editing.

1 Introduction

The use of machine translation (MT) as part of the localization workflow has mushroomed in recent years, with post-edited MT becoming an increasingly cost-effective solution for specific domains and language pairs. DePalma and Hegde (2010) stated that 42% of language service providers (LSPs) surveyed said that they offered post-edited MT to customers. At present, post-editing tends to be carried out via tools built for editing human-generated translations, such as translation memory (TM) or Translation Envi-

ronment Tools (TEnT). These environments are fairly well suited to the task for which they were intended. However, it is our opinion that integration with machine translation and support for the post-editing task are not necessarily well catered for in current translation editing interfaces. This lack of support may lead to cognitive friction during the post-editing task and to reluctance among translators to accept post-editing jobs. This paper describes the results of a survey of professional translators and post-editors, in which they chose features and functions that they would like to see in translation and post-editing user interfaces (UIs). The survey is intended as a first step towards creating specifications for UIs that better support the post-editing task. Our starting point is that translators do not require a separate editor for post-editing, but rather that features could be integrated into existing commercial tools in order to better support the task and, ultimately, integration with MT systems.

Research on post-editing has tended to focus largely on rates of productivity. Recent papers have measured translation throughput, cognitive effort, quality (as perceived when compared with human translation), or have attempted to estimate MT quality via comparison of performance with automatic evaluation metrics (AEMs) (e.g. de Almeida and O'Brien, 2010; Specia and Farzindar, 2010; Koponen et al, 2012). This research has involved the use of commercial TM tools such as SDL Trados, proprietary tools such as Crosslang, or purpose-built tools for research that have simple UIs such as Caitra (Koehn, 2009) or PET (Aziz et al., 2012). There has, however, been little focus on the UI itself, or on the functionality required for the job of post-editing.

Vieira and Specia (2011) rated several text-editing tools used for post-editing, using various

criteria, including one of “interface intuitiveness”. They acknowledge that this criterion was “highly subjective” as “its judgment was based solely on the experience of a single translator attempting to use the toolkits for the first time” (ibid.). The commercial TM tools rated all “put some effort into assigning intuitive meaning to the interface of the system” by utilizing color codes (ibid.), providing the source and target segments, including concordance search, and including dictionary and other display functions. The tools that they rated highest, however, “show clear evidence of collecting feedback from translators” (ibid.). Their wish list for a post-editing interface includes more sophisticated alignment, accurate confidence scores for MT proposals, and change tracking (included in subsequent versions of SDL Trados Studio), and they conclude that “a number of features deemed desirable for the work of a translator were not satisfactorily found in any of the tools analyzed” (ibid.).

Lagoudaki investigated text editing UIs as part of her TM survey in 2006. She found that, during development, TM users were usually “invited to provide feedback on an almost finished product with limited possibilities for changes” (2006). One translator in Moorkens (2012) said that developers had not understood her feedback as they had not worked as translators and “they don’t know the problems you encounter or the things you would like to see”. Lagoudaki also had the opinion that industry research is mostly motivated by “technical improvement of the TM system and not how the TM system can best meet the needs of its users” (2008). This runs counter to user-centered design recommendations, whereby a designer defines user profiles, usability requirements, and models before designing the UI (Redmond-Pyle and Moore, 1995).

Lagoudaki also wrote that “systems usability and end-users’ demands seem to have been of only subordinate interest” in TM system development (2008). Based on her research, the message from the users of TMs is occasionally conflicting. However, she concludes that an overall message is clear: TM users want simplicity. This does not necessarily mean fewer features; rather they want a streamlined process with compatibility between languages and scripts. They want ease of access, meaning “affordability of the system, not only in terms of purchase cost, but also

in terms of upgrade, support and training costs” (2008).

To better understand what features post-editors might require we designed a survey in which the questions focused on five areas in particular. (1) Participants were asked for some biographical details, such as years of professional experience, and about their attitude to technology. (2) They were asked about their current working methods, (3) what they would like to see in their ideal UI, (4) how they would like to see TM matches and MT output presented, and (5) about intelligent functionality that might help combine TM and MT matches. This survey is the first stage in a study that will be followed by interviews and observation, with the aim of creating specifications for a UI dedicated to the task of post-editing. Some interim results from the survey are contained in the following sections.

2 Survey Responses

The survey had 181 participants, of whom 102 completed the survey.¹ 121 participants completed the demographic section. The age range of participants is shown in Figure 1.

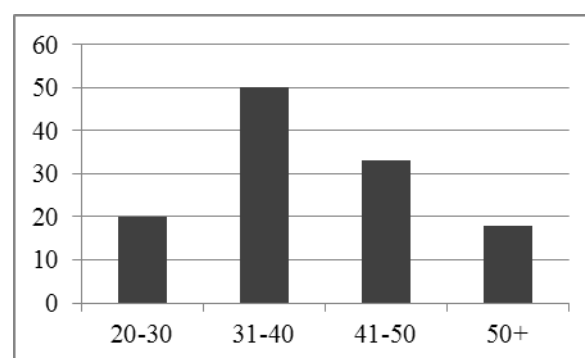


Figure 1: Age range of participants

Most participants reported that they had 8-10 years’ experience as a translator, with nine participants claiming over 20 years’ experience. Reported post-editing experience was, as may be expected, far shorter. 29 participants said that they had no experience as a post-editor, while most of the others who answered claimed 1-3 years’ experience.

¹ We are grateful to the translators who gave of their time to fill in the survey and to the following companies who promoted it: Alchemy/ TDC, Lingo24, Pactera, Roundtable, VistaTEC, and WeLocalize.

40% of participants in this section are freelancers working independently without an agency, 30% work on a freelance basis with one or more agencies, and 22% are employees of a translation or localization company. Thus, the respondents represent a good spread of work profiles. Over half of the participants (69) said that they like using TM technology, whereas only 23 (19%) said that they like using MT. 79% of participants (95) said that they use TM because it helps their work, and 36% (43) felt the same way about MT. Just over 50% (61) said that MT is still problematic.

Of 114 participants, 100 translate from English, although many listed other source languages too. The target languages are listed in Table 1. The dominance of English as a source language is probably determined by the nature of the respondents and the companies who promoted the survey via their UK or Irish offices, many of whom operate in the IT localization domain. As can be seen, there is a reasonable spread of target languages.

Target Language	No.
Arabic	2
Chinese	16
Czech	4
Danish	1
Dutch	2
English	25
Finnish	3
French	12
German	16
Greek	2
Hindi	1
Hungarian	2
Italian	6
Japanese	4
Korean	1
Malay	1
Norwegian (Bokmål)	1
Polish	1
Portuguese	12
Russian	2
Spanish	8
Swedish	3

Thai	2
Turkish	3
Urdu	2

Table 1. Participants' target languages

2.1 Current Editing Environment

107 participants provided details of the editing environment(s) that they currently use for post-editing. Most participants (76 or 70%) use more than one environment regularly. 74 (69%) use a version of the SDL Trados TM tool. There was little difference in the rate of SDL Trados use between freelancers and company employees; 21 of 28 company employees use SDL Trados. SDL Trados was also listed as the most widely-used tool in Lagoudaki (2006) with a rate of 51% usage among respondents. Lagoudaki also found that company employees are more likely to use multiple tools. In the current survey, 18 of 28 (64%) company employees said that they use multiple tools, with a similar rate of 61% (46 of 75) among freelancers. Interestingly, 44 (41%) use Microsoft Word for post-editing, which suggests that, contrary to what might be deemed best practice, MT and TM are not combined in many instances. Figure 2 shows other tools used for post-editing and the actual number of users among survey participants. Eight participants also listed proprietary tools (Translation Workspace, Helium, and MS LocStudio). Some other tools used by fewer than eight participants were Passolo (6), OmegaT (5), Star Transit (2), TransStudio (1), Alchemy Catalyst (1), and Publisher (1).

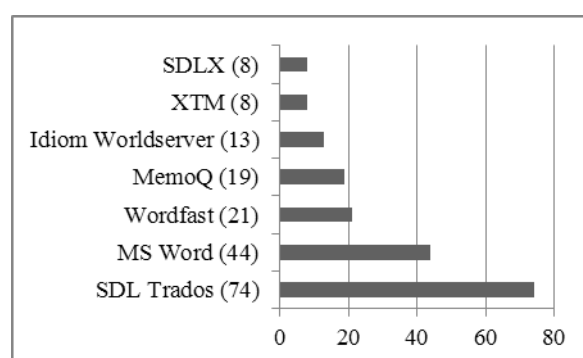


Figure 2. Tools used for post-editing

65% (80) said that they prefer to customize their editor rather than using the default set-up. Of those 80 respondents, 63 (79%) adjust their on-

screen layout, 59 (74%) adjust tag visibility, 57 (71%) adjust font type, and 19 (24%) adjust colors. Of the total number of participants who responded to the question about customization (124), roughly half are unhappy with the default layout, coloring, and display of tags in their editing tool.

2.2 UI Wish List

As might be expected from the results in the previous section, survey participants consider customizability as important. Of 119 respondents, 73 (61%) said that their ideal UI should be customizable, and 70 (59%) said that it should be clean and uncluttered. 63 (53%) always want to see the source segment alongside an MT suggestion, with 37 wishing to see them aligned horizontally and 36 preferring vertical alignment. 72 (58%) always want to see an approved glossary, although 37 (30%) would like to be able to hide or unhide the glossary.

Participants were asked what currently unavailable features they would like to see in a post-editing environment. Several suggested dynamic changes to the MT system in the case of “a recurrent MT error which needs to be fixed many times during the post-edition”. While not a trivial demand, this is an important indicator that MT and UI developers need to find efficient solutions for ‘on-the-fly’ improvements of MT output. Others requested a global find-and-replace function, dictionary plug-ins, reliable concordance features, and QA interoperability (particularly with ApSIC Xbench). What is perhaps most striking about these responses is that only one pertains specifically to post-editing and the other features can be seen as features that are desirable *in general* for supporting the translation task.

Most participants rely heavily on keyboard shortcuts. Of 119, 33% stated that they use keyboard shortcuts often, and 37% use them very often (4% never use them). 82% feel that using keyboard shortcuts improves their productivity. There are some conventions for shortcuts or text selection that users are likely to have become accustomed to. While it may be best to use these conventional shortcuts not to “reinvent too many wheels at once” (Tidwell, 2005), participants in this survey were asked whether certain shortcuts would be useful specifically for the task of MT post-editing, which can demand significant key-

boarding effort. Their responses are shown in Table 2.

Shortcut	No.
Dictionary search	103
One-click rejection of MT suggestion	96
Web-based parallel text lookup	92
Change capitalization	81
Add/delete spaces	67
Apply source punctuation to target	61

Table 2. Keyboard shortcuts requested

Again, the most popular shortcut (dictionary search) and the suggestion of a shortcut for web-based parallel text lookup are not post-editing-specific. The choice by 96 participants (81%) of a keyboard shortcut for a one-click rejection of an MT suggestion is specific to post-editing, but when taken in conjunction with the 50% who had previously said that MT is still problematic, may suggest apprehension about MT quality or usefulness among the participants. Post-editing guidelines often encourage post-editors to use as much of the MT output as possible. At the same time, segments that are completely unusable are still relatively frequent, so this one-click rejection button might actually save time. 68% suggested a keyboard shortcut for changing capitalization, recognizing that letter casing is still problematic in MT output.

Language-specific keyboard shortcut suggestions were less popular, possibly due to the large variety of target languages among participants. The most popular suggested shortcut would change the number of a word (e.g. from singular to plural), but less than half of the participants (59 of 123 or 48%) considered that such a shortcut may be useful. Further responses are shown in Table 3 with proposed shortcuts in one column and the number of respondents who said this would be useful in the other.

Shortcut	No.
Change number (sing./pl.)	59
Adjust word order	58
Change gender	48
Change verb form	46
Add/delete postposition	45
Add/delete preposition	43

Add/delete conjunction	40
------------------------	----

Table 3. Language-specific keyboard shortcuts

Participants' comments explained their misgivings relating to these shortcuts. Some were in favor of the shortcuts: "Changing gender and number of words with shortcuts would be something very useful, in my opinion." Many could not understand how they might work in practice. "Finnish is such a complicated language that those kinds of shortcuts probably wouldn't work properly," wrote one. Several thought that manual changes would be easier or less time-consuming than memorizing a large number of shortcuts. "Frankly, oftentimes it takes you less to overwrite/type what you need than learning and applying many shortcuts." Other participants did not consider the shortcuts relevant for their language pairs. "Few of the above suggested features apply to the languages I use." Participants would, however, be in favor of customizable shortcuts. 66% (81) would like to be able to add macros or scripts to adapt the UI functionality, such as adding new keyboard shortcuts. 50% (61) would like to see a guided method to help them create such a macro.

2.3 Match presentation

Most participants (93 or 79%) agreed that they would like to see MT engine confidence scores in the editing environment. Of these 93, 68 (73%) favored the presentation of such scores in the format of percentages, like a fuzzy match score in a TM tool, while 22 (24%) chose a scheme of color coding to denote confidence. Participants were asked what they would like to be shown in the case where an MT match received a higher confidence score than any fuzzy match available from the TM. 106 (90%) said that they would like to see both MT and TM matches. Apprehension about MT quality is again evident as only two participants (less than 2%) said that they would like to see the MT match only, whereas 6 said that they would still like to see the TM match only, ignoring the higher-rated MT match. This apprehension is also evident in responses to the question (shown in Figure 3): below which fuzzy match value would you prefer to see an MT rather than TM fuzzy match?

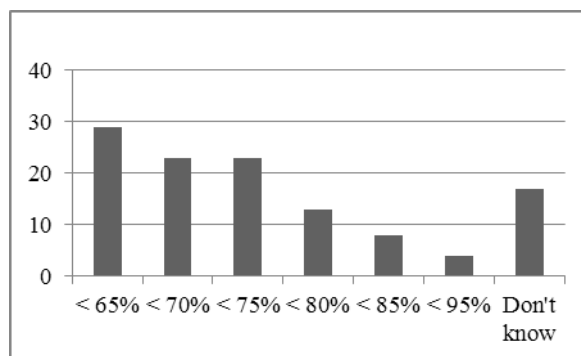


Figure 3. Fuzzy match limits

If the editing environment could combine the best sub-segments from the MT system and the TM to produce an MTM (Machine Translation/Translation Memory) match, 76 participants (64%) said that they felt it would be useful. 85 (72%) would like to see this marked as an 'MTM' match.

Comments for this question showed mixed opinions. Some participants feel positively about a potential MTM match: "Now THAT's a good idea! Somehow dynamically combine the MT output with e.g. MemoQ's longest subsegment concordance." Others are skeptical about the value of any inclusion of MT: "MT is still in baby's shoes, and the quality is horrible, so MT is not very useful in general."

Some participants commented that they would rather have the provenance of MT or TM suggestions kept separate: "It's cognitively difficult enough to distinguish between Fuzzy errors and MT errors." This was shown to be a widespread view when participants were asked whether they would like to see the provenance denoted by color or at a sub-segment level. The answer was an overwhelming 'yes' with 104 (88%) in favor. Throughout the survey, it became clear that meta-data showing the origin of match suggestions is important to translators and post-editors. Despite some misgivings about MT, 99 (84%) would like to see 'the best MT suggestion' automatically appear in the UI target window when no TM match is available.

2.4 Intelligent functionality

Participants were asked their opinion of some functions that have been suggested for post-editing of MT, such as interactive machine translation (IMT), whereby human edits are used by an MT system as "additional information to

achieve improved suggestions” (Alabau et al., 2012).

72% (83) said that, when working with a client-specific MT system, their edits should be used to improve the MT system. A further 24 (21%) were unsure, with concerns evident in the comments. Some were concerned about issues relating to confidentiality, while others resented further reuse of their translation work. This intellectual property concern was emphasized by one participant to write: “Who would pay a translator for his intellectual work in improving the TM/MTM? Generated content is usually considered property of the agency for which a freelancer works”, adding “they’ll sell you out any time, any way”.

3 Conclusion and future work

Several works have referred to TM tool users’ dissatisfaction with their current editing environments (Lagoudaki, 2008; McBride, 2009). In this paper, participants again expressed dissatisfaction with their current editing environment. What was striking here was that many comments pertained to translation editor UIs in general, which seem to still have many short-comings even before post-editing of MT output is considered. Participants emphasized the importance of customizability in their ideal UI. They would like their UI to be clean and uncluttered, and to have plugins for dictionary and Internet search, and for improved concordance search. Most participants currently use a version of SDL Trados for post-editing, and most use multiple tools. 41% use Microsoft Word, suggesting that their current workflow does not combine MT and TM.

Participants would like to see further keyboard shortcuts added to their editing environment for such functions as dictionary search, to remove an MT suggestion, or to change capitalization in their target window. However, they are more circumspect when it comes to complex, language-specific shortcuts that could change word order or gender, due to skepticism over how these functions would work in practice.

Participants have qualified enthusiasm for sub-segment integration of MT and TM, amidst concern over how well this integration might work, and over how to display the provenance of the suggested target text. While some were con-

cerned about intellectual property and confidentiality issues, they were largely in favor of dynamic improvements being made to their MT suggestions by incorporating post-edits in an IMT system.

The survey on which this paper is based is still ongoing at the time of writing. Several survey participants have chosen to waive their right to anonymity within the survey in order to participate in follow-on interviews. Based on the survey results and on these interviews, we intend to complete a specifications document for a post-editing UI, and to apply those specifications in building a new prototype UI for test purposes.

4 Acknowledgement

This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University.

References

- Alabau, Vicent, Luis A. Leiva, Daniel Ortiz-Martínez and Francisco Casacuberta. 1997. User Evaluation of Interactive Machine Translation Systems. *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT)*:20-23.
- Aziz, Wilker, Sheila C. M. de Sousa, and Lucia Specia. 2012. PET: a tool for post-editing and assessing machine translation. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 12)*, Istanbul, Turkey:20-23.
- De Almeida, Giselle and Sharon O’Brien. 2010. Analysing post-editing performance: correlations with years of translation experience. *EAMT 2010: Proceedings of the 14th Annual conference of the European Association for Machine Translation*, 27-28 May 2010, Saint-Raphaël, France. Proceedings ed. Viggo Hansen and François Yvon:8.
- DePalma, Donald A., and Vijayalaxmi Hegde. 2010. *The market for MT post-editing*. Commonsense Advisory, Boston, MA.
- Koehn, Philipp. 2009. A Web-Based Interactive Computer Aided Translation Tool. *ACL Software demonstration*.
- Koponen, Maarit, Wilker Aziz, Luciana Ramos, and Lucia Specia. 2012. Post-editing time as a measure of cognitive effort. *AMTA-2012: Workshop on*

post-editing technology and practice, San Diego, October 28, 2012:10.

- Lagoudaki, Elina. 2006. Translation Memories Survey 2006: Users' Perceptions Around TM Use. *Proceedings of ASLIB Translating and the Computer* 28. London, UK. 15-16 November 2006.
- Lagoudaki, Elina. 2008. *Expanding the Possibilities of Translation Memory Systems: From the Translator's Wishlist to the Developer's Design*. PhD thesis. Imperial College, London.
- McBride, Cheryl. 2009. *Translation Memory Systems: An Analysis of Translators' Attitudes and Opinions*. MA thesis. University of Ottawa.
- Moorkens, Joss. 2012. *Measuring Consistency in Translation Memories: A Mixed-Methods Case Study*. PhD thesis. Dublin City University.
- Redmond-Pyle, David and Alan Moore. 1995. *Graphical User Interface Design and Evaluation*. Coventry, UK: Prentice Hall.
- Specia, Lucia and Atefeh Farzindar. 2010. Estimating machine translation post-editing effort with HTER. *JEC 2010: Second joint EM+/CNGL Workshop - Bringing MT to the user: research on integrating MT in the translation industry*. AMTA 2010, Denver, Colorado, November 4, 2010:33-41.
- Tidwell, Jenifer. 2006. *Designing Interfaces*. Sebastopol, CA: O'Reilly.
- Vieira, Lucas N. and Lucia Specia. 2011. A Review of Machine Translation Tools from a Post-Editing Perspective. *Proceedings of the Third Joint EM+/CNGL Workshop - Bringing MT to the User: Research Meets Translators (JEC '11)*. Luxembourg, 14 October 2011:33-42.

Integrated Post-Editing and Translation Management for Lay User Communities

Adrian Laurenzi

Department of Computer
Science and Engineering
University of Washington
Seattle, WA, USA
alaurenz@uw.edu

Megumu Brownstein

Northwest Center of
Public Health Practice
University of Washington
Seattle, WA, USA
megumu@uw.edu

Anne M. Turner

Department of Health Services
Department of Biomedical
Informatics and Medical Education
University of Washington
Seattle, WA, USA
amtur@uw.edu

Katrin Kirchhoff

Department of
Electrical Engineering
University of Washington
Seattle, WA, USA
kk2@uw.edu

Abstract

Over the past decade machine translation has reached a high level of maturity and is now routinely utilized by a wide variety of organizations, including multinational corporations, language service providers, and governmental/non-profit organizations. However, there are many communities that could benefit greatly from machine translation but do not actively use it, either because of a lack of awareness of its current capabilities, or because of real or perceived barriers to adopting machine translation technology. In this paper we present a case study of introducing machine translation in combination with post-editing to one such community, namely employees at local and regional public health departments in the U.S. We describe a methodology for determining their translation needs, and describe the development of an integrated post-editing and translation management system specifically targeted to their typical workflow. We report results from user testing and participatory design studies and conclude with a set of recommendations and best practices for introducing machine translation plus post-editing to lay user communities.

1 Introduction

Over the past decade the development of machine translation (MT) technology has made rapid progress and has reached a high level of maturity. MT is now routinely being used for a vari-

ety of tasks. Machine translation plus post-editing (MT+PE) has been shown to significantly increase translator productivity (see e.g. (Guerberof, 2009; Plitt and Masselot, 2010; Green et al., 2013)) and has become a common procedure for many language service providers, corporations, and government organizations.

However, there are still many communities that could potentially benefit from MT but that do not currently use it. These are often non-profit, educational, faith-based, or research organizations whose members may be experts in a particular domain but who are “lay” users from an MT perspective, i.e. they are not trained translators or post-editors. Lay communities may be prevented from using MT by a lack of awareness of the current capabilities of MT systems, lack of technological know-how, costs, or by an inherent bias against MT. The adoption of MT technology would likely make the work of these communities more widely accessible, which would ultimately result in a benefit to the public. In addition it might provide unique insights or interesting data for future MT research, e.g. data from non-mainstream language pairs or specialized domains. Finally, lay communities provide a more general testbed for user modeling, adaptation, and human-computer interface design, which all need to be addressed if MT is to become a ubiquitous technology.

In order to introduce MT to lay communities it is necessary to study their current translation practices, actual translation needs, and, most importantly, to develop a model for integrating MT+PE into their typical workflows. In this paper we describe our experiences with introducing MT+PE to a community of public health professionals in lo-

cal and regional health departments in the U.S. Although our focus is on one particular community we believe that our case study can serve as a general model for lay user communities.

The paper is structured as follows: Section 2 provides the background for this study. Section 3 describes an initial data gathering phase and a pilot PE study to shed light on translation practices, translation needs, typical workflows, and the feasibility of having lay users post-edit translations of health and safety materials. Based on these initial observation we have developed an integrated post-editing and translation management system for lay user communities, described in Section 4. Section 4.2 details initial user studies and an iterative participatory design process used to refine the system. Section 5 compares this system to related work. We conclude with a summary of insights and best practices for introducing MT+PE to lay communities (Section 6).

2 Background: Machine Translation and Public Health Practice in the U.S.

The U.S. population is characterized by a fair amount of linguistic diversity. According to the 2011 American Community Survey estimates (ACS, 2011) 20.8% of the population over 5 years of age speak a language other than English at home; of these, 41.8% report speaking English “less than very well”. This percentage is even higher for certain demographic groups; e.g., it reaches 63.5% for Spanish speakers of 65 years of age or older. 24.7% to 27.7% of all households speaking Spanish or an Asian/Pacific-Island language are classified as linguistically isolated (all household members 14 years or older speak English less than “very well”). Such limited English proficiency (LEP) is correlated with adverse health outcomes. Previous studies have shown that LEP populations have more difficulty in gaining access to health care, fewer preventative health screenings, and poorer health status than English-speaking minority groups (Goel et al., 2003; Jacobs et al., 2004; Ponce et al., 2006).

This situation persists despite federal mandates requiring special provisions for LEP populations. For example, guidelines issued by the U.S. Department of Health and Human Services (DHHS) require that agencies receiving financial assistance from DHHS must take “reasonable steps”

to make their services accessible to LEP populations, which includes linguistic accessibility. In the overall healthcare context, linguistic accessibility needs to be addressed at different levels and in different forms, including providing interpreting services during patient-provider interactions, hospital discharge instructions and consent forms in different languages, or translations of newsletters or flyers on disease prevention and available health services. Here we focus on the translation and dissemination of consumer-oriented health information documents created by public health departments.

The DHHS mandate does include providing translation of vital documents (DHHS, 2003). However, in practice there is a lack of high-quality, up-to-date health information materials in languages other than English, especially at the state and local community level. The primary reasons for this situation are the lack of funds and staff time to create multilingual documents. Currently, translation practices in regional public health departments are non-standardized and vary widely. In a survey of translation practices in regional health departments in the U.S. we found that they exclusively use traditional human translation processes. Departments typically contract with a small number of language service providers. When a translation is needed, the first step is to obtain quotes from providers. Documents are then sent out to the winning bidder. The average turnaround time for translations to be completed is 15 days, with a minimum of 2 days even for rush orders. Translations then go through another internal review and quality control step before they are published. This is a time-consuming process, especially in situations where a rapid response to an emerging health crisis is required. Additionally, health departments have very scarce financial and staff resources, e.g. one medium-sized health department in Washington state reported having a monthly budget of only \$50 for translation work. Per-word translation costs reported to us by health departments participating in our study range between \$0.20 and \$1.73 (for rush orders); thus, even when using the lowest-cost service this budget allows for the translation of only 250 words per month.

MT could significantly accelerate and streamline the process of producing multilingual health information materials by eliminating the time-

consuming and costly step of outsourcing translation to external vendors. Under this model documents would first be translated automatically before being post-edited by a bilingual in-house staff member. We previously conducted a pilot study demonstrating that MT+PE leads to faster turnaround times and lower cost while the quality of the output is equivalent to human translations (?). However, our initial studies also indicated that, in order to be adopted as a standard tool, MT needs to be properly integrated into the typical workflow of public health departments and needs to be adapted to employees' needs as far as possible. Another barrier consists of attitudes and beliefs about MT. We found that employees typically are not aware of the quality of state-of-the-art MT engines, the concept of PE, or of standard support tools available.

3 Initial Feasibility Study

We conducted an initial study of human factors involved in integrating MT+PE into the standard workflow of public health departments. Our focus was on health and safety information documents (regarding e.g., vaccines, preventative screenings, infectious diseases, and maternal and child health) that are disseminated as websites, flyers, or mass mailings. We conducted 41 semi-structured interviews and 4 focus groups with health department staff involved in translation processes. The health departments included a state health department, a large municipal health department and several rural health departments that serve populations with a high percentage of LEP speakers. The participants were asked to provide a description of their current translation processes, obstacles and incentives to creating multilingual information materials, and attitudes towards MT. Transcripts of these interactions were coded using the Cognitive Workflow Analysis framework (Vicente, 1999) and the method of constant comparisons (Glaser, 1965).

The most important insights from this study can be summarized as follows:

1. Health departments typically do not have dedicated budgets or support staff for acquiring, installing and maintaining translation software, or for training employees in its use. Therefore, any MT based system targeted at health departments must be as low-cost as possible, intuitive, and easy to use, requiring as little prior technological

knowledge or initial training time as possible.

2. There is a moderate volume of translation work overall, but it may spike in response to emergencies (e.g., disease outbreaks or natural disasters) and require fast turnaround times.

3. Employees do not work on translation continuously but intermittently, in addition to other work tasks. There is a need for tracking post-editing progress, saving intermediate and partially completed work, and accessing different evolving versions of the same document (version control).

4. Different health departments could benefit greatly from sharing already-translated documents and language expertise. Currently there is no system in place to archive and share translations. Health departments often have one or two bilingual staff members but they only represent the largest language groups in the areas they serve. Thus, ideally the system should also support the online collaboration of geographically distributed workers with complementary language expertise, and the sharing of translated documents.

5. Health department employees are very concerned about the accuracy of their translated documents since they convey health and safety information. In addition translations need to be culturally appropriate and are often targeted to a specific demographic group. An MT-based system needs to have provisions for multiple layers of quality control and needs to be able to accommodate information about requirements for specific documents (e.g. desired reading level, target group, etc.).

We next conducted a pilot PE study with public-health professionals to assess whether staff members who are domain experts but not trained translators can post-edit translated documents to an acceptable standard. To this end we implemented a Java-based in-house tool that provides a simple PE interface and includes timing and keystroke logging. Source documents and their translations are displayed side-by-side in two aligned text windows, with one sentence per line. While the user is editing the translated sentences, the corresponding source sentences are highlighted. Optionally, the original non-edited machine translation can be re-displayed in a third window. Timing begins when the user first clicks inside the editable text area; it can be paused by clicking a button, which prevents the user from editing until they resume the session. All user keystrokes are logged in a single

text file per document, along with the time point at which each key was pressed. After the user has finished editing, the session is completed and the post-edited machine translation is saved as a text file along with meta-information (filename, start time, end time, pause times).

A total of 25 English health documents with an average word count of 923 (standard deviation: 452) were used for the study. Spanish translations of the documents were created using Google Translate. Eight bilingual health department staff members fluent in English and Spanish were recruited to post-edit the translations in two different sessions. They had worked in their current organization between 1.5 and 20 years and performed functions such as “Immunization Coordinator”, “Health Services Consultant”, or “Research coordinator”. Two of them had previously worked as medical interpreters; all had some experience reviewing manual translations. However, none of them were trained translators or post-editors, and they did not have experience working with machine translation or with professional translation tools. They were instructed to perform all necessary edits to create grammatically correct and accurate translations in a timely manner. Four independent quality reviewers (trained translators or interpreters recruited from health departments) were then asked to perform a blind comparison of the post-edited machine translations and human translations of the same source documents. They classified translations according to whether they were equivalent or different, and if so, which one was preferred. Post-editors were also asked about their impressions of the machine translations, which errors they found most difficult to edit, and their impressions of the PE tool.

Post-editing took on average 24.5 minutes (standard deviation: 14.9) per document. We computed the overall PE time and the average duration of pauses for each document and post-editor but did not find any consistent patterns – there was neither a strong correlation between document length and PE time nor a consistent correlation between PE time and the number of documents already completed. Some post-editors became faster from Session 1 to Session 2, indicating a learning effect, but others did not. Timing seems to be largely dependent on the individual, with some post-editors taking more time to double-check and ensure cor-

MT+PE preferred	HT preferred	Equivalent
18	16	16

Table 1: Number of votes assigned to categories in qualitative comparison of human-only translation (HT) and MT+PE output. Each of the 25 document was rated twice by two independent reviewers.

rect translations after the initial post-editing pass, whereas others do a single integrated pass over the text. This is in line with similar observations reported in the literature (O’Brien, 2006; Koponen et al., 2012).

The quality rating results (Table 1) showed that overall the quality of post-edited documents did not differ from their human-translated counterparts – preference ratings for documents that were not judged equivalent were distributed approximately evenly across the three different categories. With regard to translation errors post-editors found word order errors the most difficult to process and to correct, followed by word sense errors. They did notice a fair amount of morphological errors but these were considered less distracting. Again, this is similar to results reported in other studies (Koponen et al., 2012). They uniformly found the PE interface intuitive and easy to use and did not voice any needs for more advanced functionality.

4 Post-Editing and Translation Management System Design

In this section we describe the development of an online system designed to be integrated in the day-to-day workflow in public health departments. An initial prototype was presented in (?). Since then, we have conducted user testing with actual public health professionals and have utilized their feedback to modify the system.

4.1 System Implementation

A prototype PE and translation management system has been implemented in the form of an web-based application using the Kohana PHP framework and a MySQL database. The front-end interface was built using JQuery, Twitter Bootstrap, HTML and CSS. There are four main modules that support the main tasks of (1) uploading a document, (2) applying MT, (3) post-editing MT output, and (4) sharing and downloading the finalized translation of a document. Users

can connect to the system via any standard web browser. After registering with the system they can upload a document in the source language (English). Registration includes establishing a user profile, which include information such as the user's agency or affiliation, language expertise, and experience (such as certifications from specific professional organizations). Upon being uploaded documents are automatically translated. Our current prototype system uses the Microsoft Translator API¹, which is free for low-volume use (up to 2M characters per month) and supports up to 39 languages. However, it is in principle possible to use the Google Translate API or any other API-based translation service instead. The translated document is then added to the pool of documents in the system. Users can start post-editing a translation by "claiming" the document, which locks a document and prevents other users from post-editing it simultaneously. Once a user "unclaims" a document, other users can access it to double-check or finish the post-editing. Users can save a claimed document even if post-editing has not yet been completed and can return to it at a later time. The system tracks the progress of each document through the translation pipeline and marks each of the four stages (1) uploaded, (2) claimed, (3) post-editing in progress, and (4) completed.

For post-editing the translated documents are automatically divided into smaller chunks based on delimiters (line breaks) in the source document. For each chunk the source text is shown above the editable machine translation. When the post-editor has finished editing a block it will be saved and marked in green. Compared to an interface where the post-editing is done in a single text area this design makes it easier for post-editors to resume post-editing because all of the previously completed blocks are clearly marked. Furthermore, this design makes it easier for multiple users to contribute to the post-editing of a single document because they can clearly see which sentences still need post-editing and which ones have been completed. A screenshot of the post-editing interface for a document being translated into Spanish is shown in Figure 1. The status bar at the top of the screen is used to track a

document's progress through the pipeline. Finally the completed, post-edited documents can be downloaded.

Users can utilize the system to upload and post-edit documents created within their own health departments; they can also volunteer to post-edit documents that have been uploaded by other departments. This allows both documents and linguistic expertise to be shared across different (possibly geographically remote) health departments, which will eventually result in a more efficient utilization of resources. The interaction among different users is facilitated by a virtual discussion board included in the post-editing interface (visible in Figure 1). Users can post comments to others, e.g., in order to discuss the translation of a particular technical term.

Users can opt in to allow the source documents, original translations and post-edits to be collected for research purposes or to train customized MT models. Every document for which permission has been given is placed in an archive which can then be downloaded by the main system administrator. Thus, the system can simultaneously act as a data collection platform to collect new parallel source and target language corpora for training, updating or adapting MT models, or to create parallel corpora of machine translations and post-edits to train statistical post-editing models.

4.2 User Testing and Iterative Design

We conducted informal usability testing with six public health workers from four different programs at a state health department to obtain feedback from target users on our prototype system. Two of them were health educators, one was an environmental health investigator, one was a policy liaison, one was a graphic designer, and one was an administrative staff member.

The facilitators presented the participants with an overview of the purpose of the system and quickly demonstrated how to upload and post-edit a provided example document. For user testing the participants were separated into two teams, each of which consisted of a project manager and a bilingual post-editor who spoke English and Spanish. Each team member was assigned a task that simulated how the system would be used in practice. The tasks assigned to each team member were con-

¹<http://www.microsofttranslator.com/dev>

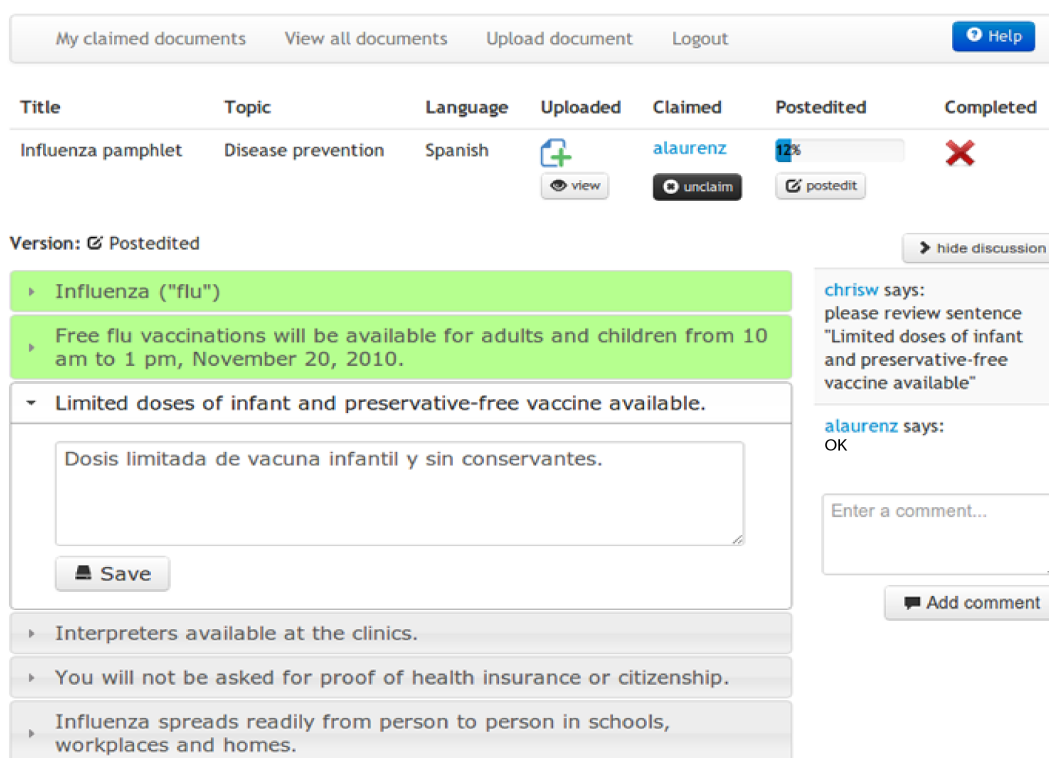


Figure 1: Screenshot of the post-editing page in our translation management system.

sistent with their roles in the translation processes used at their agency. The project manager on each team was assigned the task of uploading a provided example document. The post-editors were assigned the following tasks: (1) claim the document that was uploaded by the project manager, (2) post-edit the MT output, and (3) download the completed translation. All participants were instructed to verbalise their thought processes during the testing and were told they would not be provided with any help or guidance. During the testing the facilitators recorded general observations and critical incidences, such as difficulties encountered with the system.

After completing the assigned tasks participants were given a questionnaire to record their feedback after using the system. The questionnaire was broken up into four sections: (1) question specific to uploading a document, (2) questions specific to post-editing, (3) general questions, and (4) ranking and suggesting potential additional features. After the participants completed the questionnaires we held a focus group discussion with the participants to answer questions and record feedback that arose from the discussion. All participants successfully completed the assigned tasks without any

help from the facilitators. Based on the questionnaire responses, focus group discussion, and facilitator observations all participants were interested in using the system in their translation work and found the system to be easy to understand and use. The responses to the question “What do you like most about the system?” included:

“Fast”;
 “Ease of use and visual appearance”;
 “It’s intuitive - easy to use”; and
 “Very efficient, clear - I like the post-editing page”
 Only one participant gave feedback other than “None” to the prompt “Please describe anything that confused you while using the system.” The reason was confusion between the “Available” icon signifying a document has not yet been claimed, and the actual “Claim” button.

Based on feedback from the participants several minor adjustments and additions were made to the look and feel of the interface. Other desired features included support for more document formats (currently only plain text files and MS Word formats are supported), which will be added in the near future. The most important changes, however, involved the addition of more features that support the typical interaction of different commu-

nity members with different roles. As one participant put it: “I miss the human factor.” For example, post-editors indicated it was undesirable for a document to be automatically marked as completed after all sentences were saved. The reason is that they would like to let other staff members verify their translations and obtain feedback before finalizing a translation. In response we added a “Mark document as completed” button that appears after all sentences are saved and that needs to be explicitly clicked before advancing the document in the pipeline. In response to the project manager’s desire to communicate guidelines or notes to post-editors (e.g., desired reading level of the upload page where document-specific information can be the document, target audience, etc.) we added a field on communicated and can later be edited. Participants also expressed interest in automatic email alerts that are sent out to post-editors whose language expertise matches the desired target language of the uploaded document. Finally, they strongly advocated being able to assign ratings to, or “Like”, particular post-edits, as these would over time help to identify reliable and trusted post-editors. In sum, these are features that mirror not only the typical workflow but also the professional hierarchy or social network that exists within their community.

5 Related Work

A variety of translation management and post-editing systems have been developed in the past. Most of them (e.g. SDL Trados², Wordfast³, etc.) are commercial products aimed at language professionals, such as translators and language service providers. Their price is often prohibitive and they frequently require software installation on the user side. Other systems, such as MemSource Cloud⁴, SmartMATE (Penkale and Way, 2012), or Wordbee⁵, work in the cloud but may still be too expensive for non-professional users.

Among free or open-source systems, Google’s Translator Toolkit⁶ comes close to our requirements in that it allows collaborative post-editing and document sharing. On the other hand it lacks

essential features for our intended use, such as incorporating meta-information about documents and post-editors. A web-based translation management system intended for lay users was described in (Federmann and Eisele, 2010). However, the system solely accepts translation requests and distributes them to several back-end translation engines; there is no functionality for post-editing, version control, or user communication. Pootle⁷ is an online translation management system primarily aimed at software localization rather than document translation. Although it supports document sharing and collaboration, it requires software installation on the client side. It does support human translation and editing but does not have an integrated MT component and is thus not suitable for our purposes.

6 Conclusions

We have presented an initial feasibility study and user testing with an integrated post-editing and translation management system for delivering MT+PE technology to communities of non-professional MT users (public health professionals). Despite focusing on a single user community we believe that the insights gained from these studies apply to other, similar communities of lay users.

First, lay communities often have severely limited financial and staff resources. Software tools that support machine translation, PE, and translation management should be easily available, low-cost (ideally free), and should not require software installation and management on the client side. The user interface should be intuitive, immediately usable and should not require extensive user training.

Second, users are likely to be domain experts rather than language professionals, and they tend to work on translation on an intermittent basis in addition to other tasks. Translation tasks may be shared among different users or among different geographically distributed groups. A PE system should enable users to save and archive partially completed work, hand partial work over to other users for completion or quality control, and it should support version control.

Third, the accuracy and appropriateness of the

²<http://www.trados.com>

³<http://www.wordfast.net>

⁴<http://www.memsource.com/translation-cloud>

⁵<http://www.wordbee.com>

⁶<http://www.google.translate/toolkit>

⁷<http://www.pootle.org>

translations for particular cultural or demographic groups are often more important in scenarios such as the one described here compared to other domains. Software tools need to support meta-annotation of documents with constraints on e.g. target user group, desired reading level, etc.

Fourth, users normally do not work in isolation but in teams who place much emphasis on personal communication in their work. They often rely on the judgment of trusted users or those with a high status in their community. These intra-community social networks should be replicated in a software application by facilitating user communication via email, virtual discussion boards, etc., and by allowing users to assign ratings to specific post-editors. At the same time users would like all communication, contact information to remain private (i.e. within the organization), and they do not wish to use their personal social networking sites for work-related purposes.

The software system described above was designed to fulfill those needs; it will shortly be released as an open-source package. We hope it will be of use to other communities, e.g. non-profit organizations providing legal assistance to LEP communities. Obvious additional elements that could be integrated are a management module for domain-specific terminology lists, or translation memories. These may be added in a future version of our system.

7 Acknowledgements

This study was funded by a University of Washington Mary Gates Research Award to the first author and grant #1R01LM010811-01 from the National Library of Medicine (NLM). Its content is the sole responsibility of the authors and does not necessarily represent the view of the NLM.

References

- ACS. 2011. American community survey. <http://factfinder2.census.gov/>, downloaded June 4, 2011.
- DHHS. 2003. Department of health and human services-guidance to federal financial assistance recipients regarding title VI prohibition against national origin discrimination affecting limited english proficient persons. *Federal Register*, 68(153):47311–47323.
- Federmann, C. and A. Eisele. 2010. MT server land: An open source MT architecture. *Prague Bulletin of Mathematical Linguistics*, 94:57–66.
- Glaser, B.G. 1965. The constant comparative method of qualitative analysis. *Social Problems*, 12(4):436–445.
- Goel, M.S., C.C. Wee, E.P. McCarthy, R.B. Davis, Q. Ngo-Metzger, and R.S. Philips. 2003. Racial and ethnic disparities in cancer screening: the importance of foreign birth as a barrier to care. *J Gen Intern Med*, 18(2), pages 1028–1035.
- Green, S., J. Heer, and C. Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of ACM Human Factors in Computing Systems (CHI)*, pages 439–448.
- Guerberof, A. 2009. Productivity and quality in MT post-editing. In *Proceedings of MT Summit XII - Workshop: Beyond Translation Memories: New Tools for Translators MT*, page 8pp, Ottawa, Ontario, Canada.
- Jacobs, E.A., D.S. Shepard, J.A. Suaya, and E.L. Stone. 2004. Overcoming language barriers in health care: Costs and benefits of interpreter services. *Research and Practice*, 94(5):866–869.
- Koponen, M., L. Ramos, W. Aziz, and L. Specia. 2012. Post-editing time as a measure of cognitive effort. In *Proceedings of the AMTA Workshop on Postediting Technology and Practice*, pages 11–20.
- O’Brien, S. 2006. Pauses as indicators of cognitive effort in post-editing machine translation output. *Across Languages and Cultures*, 7(1):1–21.
- Penkale, S. and A. Way. 2012. SmartMATE: An online end-to-end MT post-editing framework. In *Proceedings of the AMTA Workshop on Postediting Technology and Practice*, pages 51–59.
- Plitt, M. and F. Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. 93:7–10.
- Ponce, N.A., R.D. Hays, and W.E. Cunningham. 2006. Linguistic disparities in health care access and health status among older adults. *Journal of General Internal Medicine*, 21:786–791.
- Vicente, K.J. 1999. *Cognitive work analysis: Towards safe, productive, and healthy computer-based work*. Lawrence Erlbaum Associates, Mahwah, NJ.

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 forums¹, 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. Guerberoof 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 tasks², 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 only³ and the others completing both bi- and monolingual tasks.

The machine translation system used in this study⁴ 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 do-

main 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 texts⁵, 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 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-

²The 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.

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

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

⁵With 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 dis-

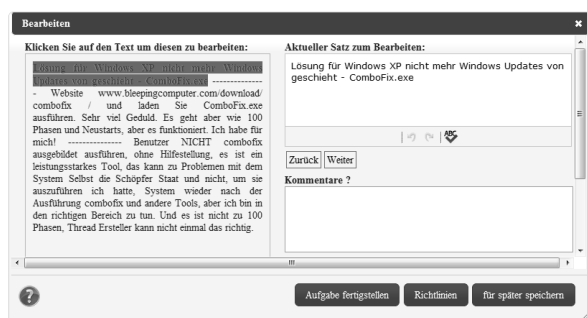


Figure 1: Post-Editing Interface

played by clicking on the "Guidelines" button:

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 languages⁶. 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

⁶The 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 %
<i>mono.</i>			
improved	67.3	57	43
same	20.4	30	29
worse	12.3	13	28
<i>bilingual</i>			
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 Ger-

man 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 %
<i>mono.</i>			
improved	63	48.6	67
same	20	25.5	18
worse	17	25.9	15
<i>bilingual</i>			
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 rank⁷ 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

⁷The 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.

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

Participant:	A	B	C	D ⁸
<i>fluency in %</i>				
improved	45	77	76	50
same	51	20	24	47
worse	4	3	0	3
<i>comprehensibility in %</i>				
improved	36	70	65	39
same	60	28	34	55
worse	4	2	1	6
<i>fidelity in %</i>				
improved	24	53	51	13
same	61	41	43	74
worse	15	6	6	13
rank (absolute)	3	1	2	4
<i>skills (Likert 1-5)</i>				
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
<i>fluency in %</i>				
improved	54	63	51	57
same	39	29	34	40
worse	7	8	15	3
<i>comprehensibility in %</i>				
improved	34	52	42	45.5
same	60	32	34	45.5
worse	6	16	24	9
<i>fidelity in %</i>				
improved	57	67	53	59
same	38	25	27	40
worse	5	8	20	1
rank (absolute)	3	1	4	2
<i>skills (Likert 1-5)</i>				
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.

⁹TER-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)

mal 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.

4.2 Monolingual vs. Bilingual

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

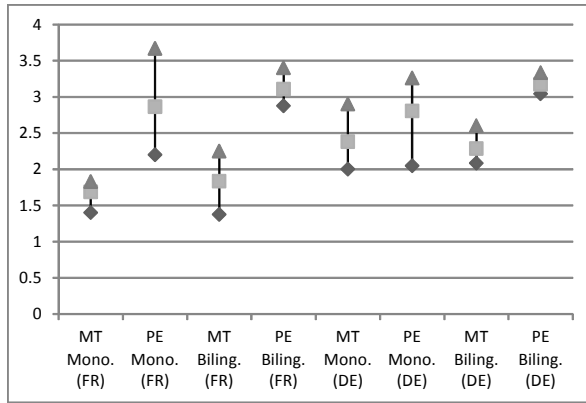


Figure 2: Fidelity Scores (Human Evaluation) with Minimum, Average and Maximum marked, with MT Mono (FR) meaning fidelity scores for the raw MT output intended for monolingual post-editing for French etc.

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.

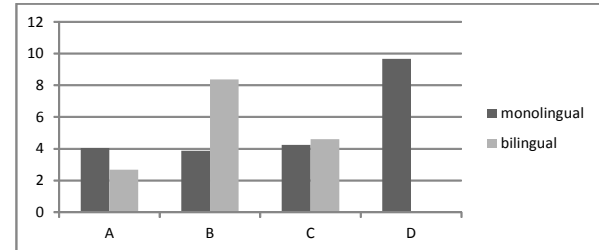


Figure 3: Time spent editing for each set-up (German) with time in average seconds per word

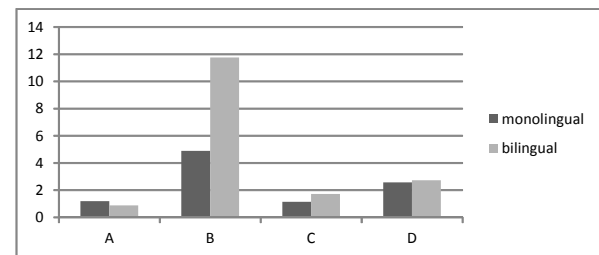


Figure 4: Time spent editing for each set-up (French) with time in average seconds per word

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. <i>If yes, how should I do.</i>	Wenn ja, wie soll ich das machen? Wenn ja, wie soll ich das tun. Wenn ja - wie soll ich es tun. <i>If yes, how should I do that?</i> Wenn ja, wie soll ich vorgehen? <i>If yes, how should I proceed?</i>
2. Thats what I was after.	Das, was ich nach. <i>This, what I after</i>	So dachte ich zumindest. <i>That's what I thought at least.</i>
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:) <i>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.</i>
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? <i>RELATED: everyone ideas, where I my wife) (or maybe have these have fun?</i>	Keine Idee wo meine Ehefrau diese Dinge her hat. <i>No idea where my wife got those things from.</i>

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 re-

trieved 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 pre-processing step.

¹⁰These 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, Ottawa, 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 Specification: 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 Munes 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)*.

Combining pre-editing and post-editing to improve SMT of user-generated content

Johanna Gerlach¹, Victoria Porro¹, Pierrette Bouillon¹, Sabine Lehmann²

(1) Université de Genève FTI/TIM - 40, bvd du Pont-d'Arve, 1211 Genève 4, Switzerland

(2) Acrolinx GmbH, Friedrichstr. 100, 10117 Berlin, Germany

Johanna.Gerlach@unige.ch, Victoria.Porro@unige.ch,

Pierrette.Bouillon@unige.ch, Sabine.Lehmann@acrolinx.com

Abstract

The poor quality of user-generated content (UGC) found in forums hinders both readability and machine-translatability. To improve these two aspects, we have developed human- and machine-oriented pre-editing rules, which correct or reformulate this content. In this paper we present the results of a study which investigates whether pre-editing rules that improve the quality of statistical machine translation (SMT) output also have a positive impact on post-editing productivity. For this study, pre-editing rules were applied to a set of French sentences extracted from a technical forum. After SMT, the post-editing temporal effort and final quality are compared for translations of the raw source and its pre-edited version. Results obtained suggest that pre-editing speeds up post-editing and that the combination of the two processes is worthy of further investigation.

1 Introduction and Background

User-generated content (UGC) such as can be found on forums, blogs and social networks is increasingly used by the online community to share technical information or to exchange problems and solutions to technical issues. Since the users contributing to the content are mainly domain specialists but not professional writers, the text quality cannot be compared with usual publishable content. In the context of a forum, where the focus is on solving problems, linguistic accuracy is often not a priority. Spelling, grammar and punctuation conventions are not always respected (cf. Figure 1). The language used is closer to spoken language, using informal syntax, colloquial vocabulary, abbreviations and tech-

nical terms (Jiang et al, 2012; Roturier and Bensadoun, 2011). Correcting or reformulating UGC is therefore not only interesting to improve readability, but also needed to improve machine-translatability.

J'ai redémarrer l'ordi (apparition de la croix rouge) mais pas besoin de restaurer le système; Toute ces mises à jour on été faite le 2013-03-13

Figure 1. Example from a forum post showing errors (agreement, word confusions) and word usage (abbreviations) typical for technical UGC

The work presented in this paper is part of the Automated Community Content Editing PorTal (ACCEPT) research project and focusses on the relationship between pre-editing and post-editing. The ACCEPT project aims at improving Statistical Machine Translation (SMT) of community content by investigating minimally-intrusive pre-editing techniques, SMT improvement methods and post-editing strategies. Within this project, the forums used are those of Symantec, one of the partners in the project. Pre-edition is carried out through the Acrolinx IQ engine and translation is done with a phrase-based Moses system.

Although several studies have explored the potential of MT of forum and user-generated content (Carrera et al, 2009; Roturier and Bensadoun, 2011; Jiang et al, 2012), few of them have looked into the role of pre- and post-editing as MT complementary modules (Aikawa et al, 2007).

In previous work (Gerlach et al., 2013), we have shown that it is possible to develop pre-editing rules that significantly improve MT output quality, where improvement was assessed through comparative evaluation. In this paper we intend to investigate whether pre-editing rules that have a positive impact on the raw SMT out-

put also have an impact on post-editing temporal effort, which is generally considered one of the most important factors in post-editing evaluations (Krings, 2001). It could be that even though the quality of raw MT output is improved, this does not facilitate the post-editor's task. We will also compare the time required for pre-editing and post-editing tasks and investigate whether time can be gained by combining both activities. Furthermore, we will analyse the final translation quality and look at the satisfaction of the post-editors.

Our aim in this study is twofold, namely: 1) ascertain whether pre-editing rules that improve MT can reduce post-editing effort, and 2) confirm that comparative human evaluation is a valid method to evaluate and select such rules, thus justifying the use of this evaluation method for the ACCEPT project.

In the next sections (2 and 3), we briefly describe the pre-editing approach used in the ACCEPT project. In section 3 we describe the experimental setup and the methodology followed. The data obtained for each experiment is analysed in section 4. Conclusions and future work are presented in section 5.

2 Pre-edition in ACCEPT

In ACCEPT, pre-edition is carried out through the Acrolinx IQ engine, which supports spelling, grammar, style and terminology checking (Brendenkamp et al, 2000). This rule-based engine follows a phenomena-oriented approach to language checking, using a combination of NLP components such as a morphological analyser and a POS tagger to obtain linguistic annotations which can be used to define complex linguistic objects. These are then used in declarative rules written in a formalism similar to regular expressions that marks phenomena that should be pre-edited. Rules can also include correction suggestions, making the pre-editing process semi-automatic, where users only have to accept suggestions provided by the system.

The Symantec community will have access to the Acrolinx engine through a browser plugin, allowing the users to check their text and apply the rules directly in the browser window when writing a forum post (Accept Deliverable D5.2, 2013). The interface of the pre-editing plugin is shown in Figure 2.

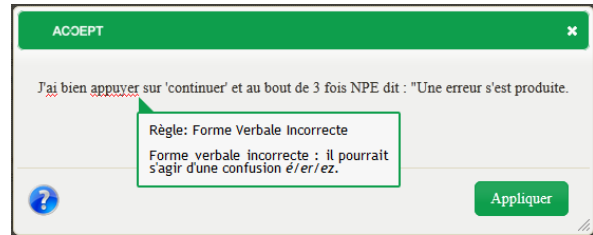


Figure 2. ACCEPT pre-editing plugin. Example of a rule which detects incorrect verb forms.

2.1 Pre-editing rules

During the first period of the project, a stable set of rules with significant positive impact was developed from scratch for French technical UGC. The rules focus mainly on four phenomena, which were proven troublesome for SMT: word confusion (due to homophones), informal and familiar French, punctuation, and structural divergences between French and English. The main criteria for their definition have been precision and impact on translation into English. Impact on translation has been assessed through human comparative evaluation, performed by advanced translation students as well as Amazon Mechanical Turk judges (Gerlach et al., 2013).

The rules are grouped into three sets. Besides the obvious separation of rules for humans and rules for the machine (Hujisen, 1998), they are grouped according to the pre-editing effort they require. Indeed, considering the end-users of the rules, namely forum users who might not be inclined to invest much time in pre-edition, we intended to offer several pre-editing options that would require different amount of involvement.

Some of the rules treat unambiguous cases and have unique suggestions. These are therefore grouped in a set (Set 1) which can be applied automatically with no human intervention. This contains rules for homophones, word confusion, tense confusion, elision and punctuation. Examples are shown in Table 1.

	Raw	Pre-edited
Source	oups j'ai oublié , j'ai sa aussi.	oups j'ai oublié, j'ai ça aussi.
SMT output	Oops I forgot, I have its also.	I have forgotten, I have this too.
Source	avez vous des explications ou astuces pour que cela fonctionne?	Avez-vous des explications ou astuces pour que cela fonctionne?
SMT output	Have you explanations or tips for it to work?	Do you have any explanations or tips for it to work?

Table 1. Examples for rule set 1

The remainder of the rules for humans have either multiple suggestions or no suggestions, thus requiring human intervention. These are grouped in a second set (Set 2), which contains rules for agreement (subject-verb, noun phrase, verb form) and style (cleft sentences, direct questions, use of present participle, incomplete negation, abbreviations), mainly for correcting informal/familiar language. An example is shown in Table 2.

	Raw	Pre-edited
Source	Tu as lu le tuto sur le forum?	As-tu lu le tutoriel sur le forum?
SMT output	You have read the Tuto on the forum?	Have you read the tutorial on the forum?

Table 2. Example for rule set 2

Finally, a third set (Set 3) contains the rules for the machine that should not be visible to end-users. The rules in this set modify word order and frequent badly translated words or expressions to produce variants better suited to MT. One important rule converts the informal second person (*Tu as compilé?*) into its formal correspondent (*Vous avez compilé?*), more frequent in the training data (Rayner et al, 2012). Another rule deals with French clitics that are easily confused with definite articles, replacing them with less ambiguous structures. Examples are shown in Table 3.

	Raw	Pre-edited
Source	J'ai apporté une modification dans le titre de ton sujet.	J'ai apporté une modification dans le titre de votre sujet
SMT output	I have made a change in the title of tone subject	I have made a change in the title of your issue
Source	Il est recommandé de la tester sur une machine dédiée.	Il est recommandé de tester ça sur une machine dédiée.
SMT output	It is recommended to the test on a dedicated machine.	It is recommended to test it on a dedicated machine.

Table 3. Example for rule set 3

In the rest of the paper we describe the experimental setup with the different tasks, the evaluation methodology and the results obtained.

3 Experiment Setup and Methodology

3.1 Corpus

The data used for this study is extracted from the French Symantec forums, where users discuss technical problems with anti-virus and other security software.

In order to create a representative corpus, we selected 684 sentences from the data provided by Symantec, based on bigram frequency, keeping the same proportion of sentences of each length. Sentence lengths range from 6 to 35 words. As a result of this selection process, all sentences were out of context.

Due to the characteristics of UGC, the segmentation of forum data into sentences is not always straightforward. Consequently, some of the automatically extracted sentences are in fact only fragments of the sentences as intended by their authors and can be difficult to understand out of context. We chose not to remove these at this stage, as we did not want to alter the data.

3.2 Participants

For both the pre-editing and post-editing tasks, we recruited translation students in the second year of the MA program at the Faculty of Translation and Interpreting (FTI) of the University of Geneva. For the pre-editing task, we recruited a native French speaker. For the post-editing task, we recruited three native English speakers who had French as a working language. None of the participants had any specific technical knowledge.

3.3 Pre-editing Task

The pre-editing task was divided in three steps. First, we applied the rules from Set 1 automatically, using Acrolinx's AutoApply Client, which replaces each flag (marked phenomena) with the first suggestion available. Since the precision of the rules is not perfect, this step can induce minor deterioration of some sentences, which we did not correct. In a second step, we had the French translator manually apply the rules from Set 2 using Acrolinx's MSWord plugin. This plugin marks all incorrect words in colour, provides information about the error in a contextual menu and, if suggestions are available, allows the user to select a correction from a list. The translator also corrected spelling errors flagged by the Acrolinx spelling module. The pre-editor was asked to treat all correct flags. During this process, we logged the keystrokes, mouse clicks and time. In a third step, we applied Set 3 automatically, using the same method as for Set 1. 456 of the original 684 sentences were affected by pre-editing, i.e. had one or more changes. The flags reported at each step are summarized in Table 4.

Set	grammar, punctuation	style, reformulations	spelling
1	87	7	-
2	74	115	362
3	-	191	-
total	161	313	362

Table 4. Flags for each step

3.4 Translation and Data Selection

The 456 sentences affected by pre-edition were then translated into English using the project's baseline system, a phrase-based Moses system, trained on translation memory data supplied by Symantec, europarl and news-commentary (Accept Deliverable D4.1, 2012).

For 319 sentences, the translation of the pre-edited version was different from that of the raw version.

In order to retain only those sentences where pre-edition had a positive impact on MT output, the translation results (319) were submitted to a comparative evaluation, on the same principle as what was done in previous works (Gerlach et al, 2013). This evaluation was performed by three bilingual judges, using a five-point scale {raw better, raw slightly better, about equal, pre-edited

slightly better, pre-edited better}. The "better" and "slightly better" judgments for each category (raw and pre-edited) were regrouped and the majority judgement for each sentence pair was calculated. The results of the comparative evaluations are shown in Table 5. When considering the majority judgements, the pre-editing rules have a significant positive impact on translation quality. In 65% of cases, translation was improved, while degradation was only observed in 11% of cases. For this specific work, we only considered unanimous judgements. Only those sentences where all three judges considered that pre-editing had had a positive impact on the translation were retained for the post-editing task. This selection had the additional benefit of removing problematic sentences, as we had noticed that judges often fail to reach a unanimous judgement when the presented sentences are difficult to understand, due to bad segmentation or very poor language quality. This final selection resulted in a set of 158 sentences, which added up to 2524 words.

Total sentences	Raw better	About equal	Pre-edited better	No majority judgement	p-value
Majority judgements					
319	34 (11%)	63 (20%)	209 (65%)	13 (4%)	<0.0001
Unanimous judgements					
193	11 (6%)	24 (12%)	158 (82%)	-	<0.0001

Table 5. Comparative evaluation

3.5 Post-editing Task

The resulting set of 158 sentences was used to investigate bilingual post-editing productivity as well as the impact of pre-edition on the quality of the final output after post-editing. Translators were asked to post-edit the machine translation output both of the raw source and of its pre-edited counterpart. This added up to a total of 316 sentences, which were randomly distributed in 71 sets of 20 pairs each.

The post-editing task was performed using the project's post-editing portal (<http://www.accept-portal.eu>, Accept Deliverable D5.2, 2013; cf. Figure 3). The portal logs editing time as well as

keystrokes for each source-target pair. This data can be exported in XLIFF format.

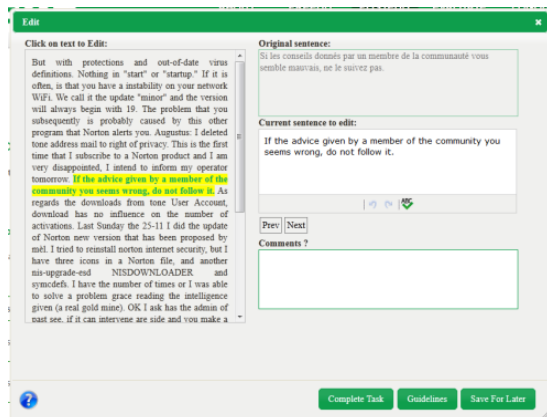


Figure 3. Post-editing Portal Interface

Post-editors were presented with a source-target pair, where the target was the machine translation of either the raw or the pre-edited sentence. Post-editing guidelines and a glossary for the domain covered by the data were provided. Post-editors were asked to render a grammatically correct target sentence, which should convey the same meaning as the original, while using as much of the raw MT output as possible. Terminology and style were not given priority. No time limit was given and all participants were paid.

At the end of the task, the participants were asked to complete a short questionnaire, which was designed to gather information about the post-editors' profile, their previous experience with MT and post-editing, and their feelings towards it.

In this experimental setup, post-editors processed each sentence twice: once the translation of the raw source and once the translation of its pre-edited counterpart. As the sentences were presented in a random order, in some cases the translation of the raw source was treated before that of the pre-edited source and vice-versa. It is logical to expect the post-editor to spend more time reading and post-editing the first instance of a pair of sentences. When the second instance appears, the post-editor has at least already read and processed the meaning of the source and thus will probably spend less time in post-editing. Since the order randomisation of our data produced an unfair distribution (69 pre-edited first vs 89 raw first), we chose to remove 20 sentences where the translation of the raw source had been processed in the first place, in order to balance the impact of processing order.

The quality of the final translations was evaluated using the LISA QA Model. The errors in all 276 sentences for each of the three post-editors were annotated by two bilingual persons, whose annotations were then put in common and discussed to resolve ambiguities and disagreements.

In the next section, we will present the results for all tasks.

4 Results

4.1 Pre-editing Effort

The pre-editor spent 53 minutes processing the entire corpus (684 sentences) using the MSWord Plugin, making 334 keystrokes, 576 left-clicks and 542 right-clicks. This process changed 567 tokens in the corpus and affected 456 sentences (cf. Table 6).

The pre-editor found the rules straightforward to apply and the pre-editing process globally quite easy, except for some terminology issues related to the unfamiliar domain.

Pre-editing task : 456 sentences	
Total time (mins)	53
Total keys	334
Total mouse-clicks	1118

Table 6. Pre-editing effort

4.2 Post-editing Effort

The post-editing effort in terms of time and keystrokes is clearly lower for the translations of pre-edited sentences. While the post-editing speed differs strongly among post-editors, the relative time gain is very similar for all three. On average, the total post-editing time for all 138 sentences is reduced by 47% with $sd=4\%$. The one-tailed t-test shows that the difference is highly significant for all three post-editors ($p<0.0025$, $t=4.581/3.094/3.635$). The results for the three post-editors are shown in Table 7.

Post-editing task : 2*138 sentences (2*2194 words)						
	PE 1		PE 2		PE 3	
	Raw	Pre-edited	Raw	Pre-edited	Raw	Pre-edited
Total time (mins)	53	29	98	56	109	54
Total keys	3492	1763	3907	2181	5579	3263
Processing speed (w/mins)	41	76	22	39	20	40

Table 7. Pre- and post-editing effort

Table 8 shows an example of a sentence before and after pre-editing, with its corresponding MT output and the post-editing times for each post-editor (in seconds).

	Raw	Pre-edited
Source	quelqu'un a t'il déjà rencontré se problème?.....	quelqu'un a-t-il déjà rencontré ce problème?.....
SMT output	Someone has it already you encountered is problem?.....	Has anyone had this problem?.....
Post-editing time (PE1/PE2/PE3)	7.7s/14.2s/16.1s	5.1s/0s/6.5s

Table 8. Examples of MT output with corresponding post-editing time

The histogram in Figure 4 illustrates the results presented in table 7. It represents the frequency distribution of time gain percentages from raw to pre-edited for each of the post-editors, which were calculated per sentence, in relation to the time used to post-edit MT output of the raw sentence. The data range is distributed into “bins” of equal size on the x axis and the frequency within each bin is shown vertically on the y axis. 25 outliers¹ were removed.

Although the post-editing time for the pre-edited sentence is not always lower than the time for the raw sentence, we observe that the cases where pre-editing reduces the post-editing time are more frequent. 312 of 389 sentences are plotted on the right-hand side of the histogram.

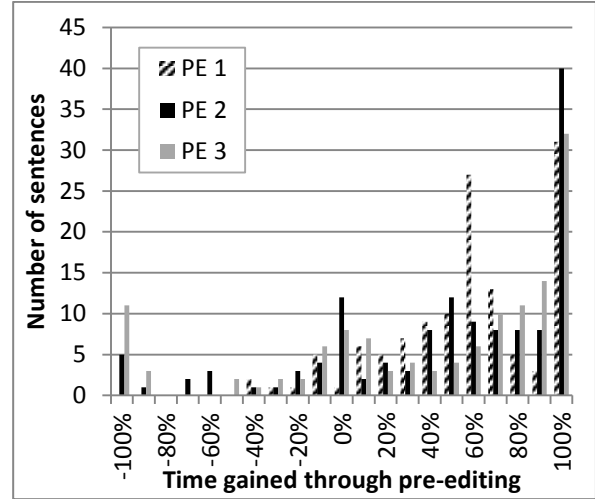


Figure 4. Distribution of relative time gained for each post-editor

While the absolute pre- and post-editing times may not be directly comparable, due to the different number of sentences processed and to the possibly artificially low post-editing times caused by the double processing mentioned in section 3.5, it remains interesting to combine these times. As not all pre-edited sentences have been post-edited, we have estimated the pre-edition time for the effectively post-edited sentences proportionally to the number of sentences, based on the data shown in Table 6, resulting in an approximate pre-editing time of 16 minutes for 138 sentences.

We observe that, for our set of sentences where pre-editing had a positive impact on MT output, the post-editing time gained by using a pre-edited source (respectively 24/42/55 minutes for each of the post-editors, cf. Table 7) outweighs the time invested in the pre-editing process itself. Combined results are shown in Table 9. Furthermore, it can be argued that for an equal time investment, the pre-editing effort is “cheaper” than the post-editing effort, as 1) it is a monolingual process, thus requiring less qualification from the user, and 2) it is semi-automatic, as most of the rules have suggestions and can be applied by selecting an item in a list.

¹ We apply one of the common definitions of outliers using the interquartile range (IQR): lower than the 1st quartile minus 1.5*IQR or greater than the 3rd quartile plus 1.5*IQR.

Combined time for 138 sentences (mins)						
	PE1		PE2		PE3	
	Raw	Pre-edited	Raw	Pre-edited	Raw	Pre-edited
Pre-editing	-	16	-	16	-	16
Post-editing	53	29	98	56	109	54
Total	53	46	98	81	139	78

Table 9. Combined pre- and post-editing times

As another indicator of post-editing effort in terms of number of edit operations, we computed the Translation Error Rate (TER) (Snover et al., 2006) for each of the two MT outputs (raw and pre-edited) using the three corresponding post-edited versions as reference. The case sensitive TER score for the translation of the raw source is 20.17, the score for the translation of the pre-edited source is 10.76, indicating a lower number of edits for the pre-edited version.

4.3 Quality Evaluation

On the whole, pre-edition seems to slightly reduce the number of errors in the final output, but the number of errors is insufficient to determine whether the difference is significant (cf. Table 10). A similar number of errors was found for all three post-editors in both versions, although far less time was spent post-editing the pre-edited version. We can therefore assume that the increase in processing speed does not entail an increase in the number of errors.

Total errors			
	Raw	Pre-edited	Reduction
PE 1	44	37	7
PE 2	28	29	-1
PE 3	41	35	6

Table 10. Error counts for each post-editor

A closer examination of the individual annotated errors does not indicate a clear relation between the errors and the output that was post-edited (MT of raw sentence or MT of pre-edited sentence). However, we have observed that there are proportionally more sentences with errors among those with longer edit distances (Levenshtein) between the raw MT output and the post-edited version. This supports the assumption that post-editors will make fewer errors when presented with a relatively clean MT output needing only few edits (rather than an output that requires heavy reformulation and corrections at

many places). While our data is insufficient to quantify this claim, this observation suggests that pre-editing can also have a positive impact on final post-edited translation quality.

Table 11 shows the error counts by category, averaged over the three post-editors. Mistranslations are the most frequent type of error, which was to be expected considering that 1) the sentences were out of context and sometimes badly segmented, making them difficult to understand, 2) the post-editors were not familiar with the domain, 3) the post-editors, not being native French speakers, might have had difficulties understanding the colloquial French used on the forums. The only category where we observe no improvement is terminology, but the number of errors is too small to be significant. The most important reduction can be observed for language errors, which include spelling, punctuation, grammar and semantics.

Final Quality Evaluation (LISA QA)			
Average per category	Raw	Pre-edited	% error reduction
Mistranslation	17.3	16.7	-4%
Accuracy	6.0	5.7	-6%
Terminology	1.3	2.3	75%
Language	9.7	6.0	-38%
Style	3.3	3.0	-10%
TOTAL	37.7	33.7	-11%

Table 11. Average error counts by error category

Most of the errors observed in our data can be attributed to typos, lack of attention and hesitation to seriously reformulate the MT output, which can at least partially be explained by the participants profiles and insights described in the next section.

4.4 Questionnaire. Insights from participants.

After the post-editing task, we asked participants to complete an anonymous questionnaire to establish their profiles and gather their insights about the post-editing task. This questionnaire was based on the questionnaire used in another experiment performed at FTI, also involving translation students, texts from the same forum and the same MT system (Morado Vázquez et al., 2013), where globally feedback was very positive. From the analysis of the answers provided,

we gathered the following information. All participants claimed to translate about 250 words per hour on an average 8-hour day of work, but had little experience as professional translators (only one claimed to have been working as a freelance for 2 years) and had hardly ever post-edited MT-output before. As for CAT tools, one only uses them when required to do so and the other two have tried them but do not use them on a daily basis.

Participants were not familiar with the topic or with Symantec products. Two found the task difficult from a terminology point of view and one indicated she had mainly experienced linguistic-related doubts.

More interestingly, when asked about the helpfulness of MT proposals to produce a final translation, two seemed sceptical (they responded 3 on a 6-point scale, where 6 stood for “Not at all, I would have preferred working from scratch”) and the third was negative (she responded 5). Nonetheless, we observed that their attitude towards post-editing itself was quite positive: they considered that post-editing was “definitely needed [...] and can help a lot” (PE1) and “useful” (PE2), except for the third participant, who found post-editing harder than translating from scratch. Despite this, they all agreed in saying that if more context was provided and if they mastered the domain or topic of the texts, they would find post-editing machine translations more useful and interesting.

5 Conclusion and Future Work

We have observed that pre-editing rules that have a significant positive impact on translation output also have a significant positive impact on post-editing time, reducing it almost by half. The combination of pre-editing and post-editing to process user-generated content seems promising, as easy monolingual pre-editing effort effectively reduces the more tedious bilingual post-editing effort. Based on the fact that a translation judged as being better is also faster to post-edit, we conclude that comparative evaluation is a valid method to select pre-editing rules for a workflow such as envisaged in the ACCEPT project. We plan to extend our investigations to examine whether pre-editing that does not directly improve translation quality also has an impact on post-editing effort.

While pre-editing does not significantly improve the quality of the final post-edited transla-

tions, there is no loss of quality linked to the time gain. The most frequent errors in the final translations are mistranslations. While the bad segmentation and lack of context are probably responsible for many of these, we suspect that the lack of experience and insufficient domain knowledge of the MA students have also influenced the results. In order to refine these results, we plan to perform in-context tests, processing entire forum posts, using both professional translators and savvy real users. This would give us more information about the causes of the mistranslations and might point to phenomena that could be corrected by pre-editing.

Finally, regarding the pre-editing task, we would like to see how pre-editors apply the rules, i.e. if, in non-controlled circumstances, they will apply all rules systematically or choose only those they consider useful.

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 D4.1 (2012), <http://www.accept.unige.ch/Products/>
- Accept Deliverable D5.2 (2013), <http://www.accept.unige.ch/Products/>
- Aikawa, Takako, Schwartz, Lee, King, Ronit, Corston-Oliver, Mo, Lozano, Carmen. 2007. Impact of Controlled Language on Translation Quality and Post-editing in a Statistical Machine Translation Environment. In *Proceedings of the MT Summit XI*, 10-14 September, Copenhagen, Denmark, pp.1-7.
- Allen, Jeffrey. 2003. “Post-editing”, in Somers, Harold: *Computers and Translation. A Translator's Guide*, John Benjamins Publishing Company, Amsterdam/Philadelphia, p. 297-317.
- Bredenkamp, A., B. Cysmann and M. Petrea. 2000. Looking for errors: A declarative formalism for resource-adaptive language checking. In *Proceedings of LREC*. Athens, Greece.
- Carrera, Jordi, Olga Beregovaya and Alex Yanishevsky. 2009. Machine Translation for Cross-Language Social Media, available: http://www.promt.com/company/technology/pdf/machine_translation_for_cross_language_social_media.pdf [accessed May 23rd 2013]

- Gerlach, Johanna, Victoria Porro and Pierrette Bouillon. 2013. La prédiction avec des règles peu coûteuses, utile pour la TA statistique des forums ? In *Proceedings of TALN/RECITAL 2013*. Sables d'Olonne, France.
- Hujisen, W. O. 1998. Controlled Language: An introduction. In *Proceedings of CLAW 98* (pp. 1–15). Pittsburg, Pennsylvania: Language Technologies Institute, Carnegie Mellon University
- Jiang, Jie, Andy Way and Rejwanul Haque. 2012. Translating User-Generated Content in the Social Networking Space. In *Proceedings of AMTA 2012*, San Diego, CA, United States.
- Krings, Hans P. 2001. Repairing texts: Empirical investigations of machine translation post-editing process. The Kent State University Press, Kent, OH.
- Morado Vázquez, Lucía, Silvia Rodríguez Vázquez and Pierrette Bouillon. 2013. Comparing forum data post-editing performance using translation memory and machine translation output: a pilot study. In *Proceedings of 14th Machine Translation Summit*, 2013, Nice, France.
- O'Brien, Sharon and Johann Roturier. 2007. How Portable are Controlled Languages Rules? A Comparison of Two Empirical MT Studies. In *Proceedings of the MT Summit XI*, Copenhagen, pages 105–114.
- Rayner, Manny, Pierrette Bouillon and Barry Haddow. 2012. Using Source-Language Transformations to Address Register Mismatches in SMT. In *Proceedings of AMTA*, San Diego, CA, United States.
- Roturier, Johann, and Anthony Bensadoun. 2011. Evaluation of MT Systems to Translate User Generated Content. In *Proceedings of the Thirteenth Machine Translation Summit*, 244–251.
- Snover, M., B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul. 2006. A Study 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.

Advanced Computer Aided Translation with a Web-Based Workbench

Vicent Alabau^{*}, Ragnar Bonk[†], Christian Buck[‡], Michael Carl[†], Francisco Casacuberta^{*}

Mercedes García-Martínez[†], Jesús González^{*}, Philipp Koehn[‡], Luis Leiva^{*}

Bartolomé Mesa-Lao[†], Daniel Ortiz^{*}, Herve Saint-Amand[‡], Germán Sanchis^{*}, and Chara Tsoukala[‡]

^{*}Institut Tecnològic d'Informàtica, Universitat Politècnica de València, Spain

[†]Copenhagen Business School, Department of International Business Communication, Denmark

[‡]School of Informatics, University of Edinburgh, Scotland

valabau@iti.upv.es, ragnar.bonk@gmx.de, cbuck@lantis.de, mc.isv@cbs.dk, fcn@iti.upv.es

mgarcia@iti.upv.es, jegonzalez@dsic.upv.es, pkoehn@inf.ed.ac.uk, luileito@iti.upv.es

bm.ihc@cbs.dk, dortiz@iti.upv.es, hsamand@inf.ed.ac.uk, gsanchis@dsic.upv.es, x.tsoukala@gmail.com

Abstract

We describe a web-based workbench that offers advanced computer aided translation (CAT) functionality: post-editing machine translation (MT), interactive translation prediction (ITP), visualization of word alignment, extensive logging with replay mode, integration with eye trackers and e-pen. It is available open source and integrates with multiple MT systems.

The goal of the CASMACAT project¹ is to develop an advanced computer aided translation workbench. At the mid-point of the 3-year project, we release this tool as open source software. It already includes a wide range of novel advanced types of assistance and other functionalities that do not exist together in any other computer aided translation tool.

The CASMACAT is working in close collaboration with the MATECAT project², which also has the goal of developing a new open source web-based computer aided translation tool, and focuses mainly on post-editing machine translation, adaptation methods, and ease of use that make such a tool suitable for professional users.

Through this combined effort, we hope to kick-start broader research into computer aided translation methods, facilitating diverse translation process studies, and reach volunteer and professional translators without advanced technical skills.

The tool is developed as a web-based platform using HTML5 and Javascript in the Browser and PHP in the backend, supported by a CAT and MT server that run as independent process (both implemented in Python but integrating tools written in various other programming languages).

¹<http://www.casmacat.eu/>

1 Related Work

There is increasing evidence for productivity gains of professional translators when they post-edit machine translation output.

For instance, Plitt and Masselot (2010) compare post-editing machine translation against unassisted translation in a web-based tool for a number of language pairs, showing productivity gains of up to 80%. Skadiņš et al. (2011) show a 30 percent increase for English-Latvian translation with a slight but acceptable degradation in quality. Federico et al. (2012) assess the benefit of offering machine translation output in addition to translation memory matches (marked as such) in a realistic work environment for translators working on legal and information technology documents. They observe productivity gains of 20-50%, roughly independent from the original translator speed and segment length, but with different results for different language pairs and domains. Moreover, Pouliquen et al. (2011) show that, aided by machine translation, non-professional post-editors may be able to create high-quality translations, comparable to a professional translation agency.

So far, usage of machine translation technology has concentrated on human-computer interaction involving the human translator as a post-editor, but rarely involves the human translator influencing the decisions of the machine translation system. Recent efforts on building interactive machine translation systems include work by Langlais et al. (2000) and Barrachina et al. (2009). Both studies develop research systems looking into a tighter integration of human translators in MT processes by developing a prediction model that interactively suggests translations to the human translator as he or she types. Related work displays several word and phrase translation choices to human translators (Koehn, 2010).

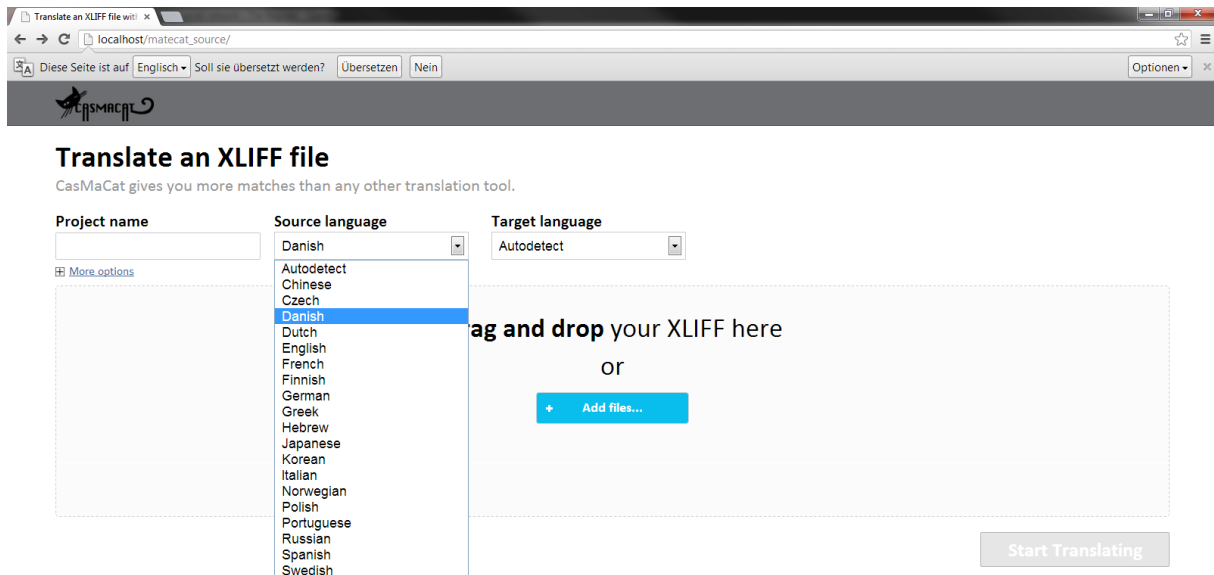


Figure 1: View for uploading new documents

2 User Interface

The CSMACAT UI consists of views designated for different tasks. The translate view is its central view, where the user can translate a document and post-editing assistance and logging takes place. Other views offer a way to upload new documents or to manage the documents that are already in the system. Also, a replay mode has been implemented. The different views will now be shown and described in the sequence they are typically used.

2.1 Upload

If the user opens the default URL without giving any special parameters, he or she is taken to the upload view. This is currently the entry point of the application. See Figure 1 for a screenshot. At this point, a user can specify one or several documents to upload and to translate. The documents uploaded must be in XLIFF format. The language pair can either be chosen manually or auto-detected from the XLIFF file. If several documents are uploaded at once, they are bundled into one job and are translated in a sequence. If the user clicks on the *Start Translating* button he or she is taken to the translate view and can start working.

2.2 Post-Editing

In the translate view, the user can now translate the document (see Figure 2). The document is presented in segments, while the currently active seg-

ment is highlighted and assistance is provided for this segment. If using the post-editing configuration without ITP up to three MT or TM suggestions are provided, from which the user can choose. The user can use shortcuts, for instance, to go to the next segment or to copy the source text to the target. The user can assign different status to a segment, for instance, *translated* for finished ones or *draft* for segments, where he or she is not yet sure about the translation and he or she wants to review later. When finished, the *Download Project* button may be used to download the translated document, again in the XLIFF format.

When in the translate view, all the actions of the user that are related to the translation task, e.g. typing, choosing a suggestion, closing a segment and so on, are logged by the CSMACAT logging module. In addition to traditional key and mouse logging, we also provide text change logging based on the HTML5 *input* element. This makes the log of text activities much more robust, e.g. it allows to log changes from paste or cut actions triggered by the browser's menu bar or the context menu of the mouse. Mouse clicks are still logged to track user interactions with UI elements. Key logging is helpful for offline analysis.

2.3 Interactive Translation Prediction

In the following paragraphs we present a short description of the main advanced CAT features that we implemented in the workbench. Such features

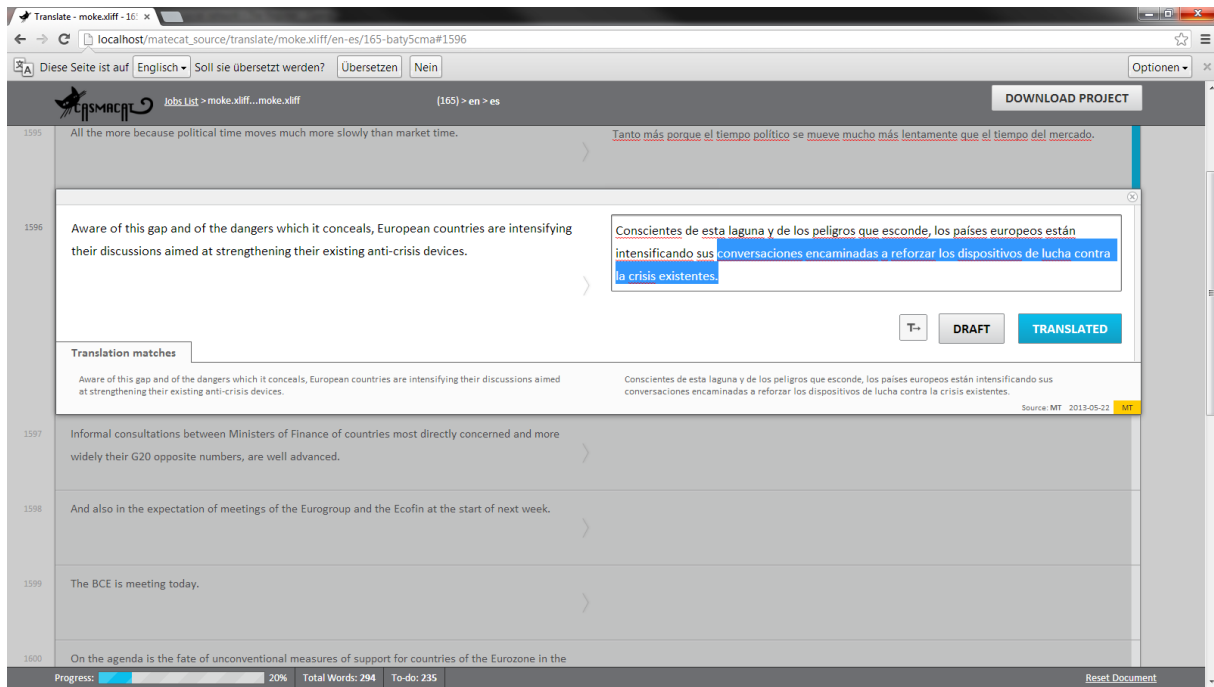


Figure 2: Translate view with post-editing configuration

are different in nature, but all of them aim at boosting translator productivity.

Intelligent Autocompletion Interactive translation prediction takes place every time a keystroke is detected by the system (Barrachina et al., 2009). In such event, the system produces a prediction for the rest of the sentence according to the text that the user has already entered. This prediction is placed at the right of the text cursor.

Confidence Measures Confidence measures (CMs) have two main applications in MT (González-Rubio et al., 2010). Firstly, CMs allow the user to spot wrong translations (for instance, by painting in red those translations with very low confidence). Secondly, CMs can also inform the user about the translated words that are possibly incorrect, but still have a chance of being correct (for instance, painted in orange). In our workbench, both applications are handled by means of two thresholds, one that favors precision and another that favors recall of changes to highlighted words.

We use confidence measures to inform the user about translation reliability under two different criteria. On the one hand, we highlight in red color those translated words that are likely to be incorrect. We use a threshold that favors precision in

Lisboa y Madrid **desee embarcarse en** un camino **diferente del adoptado** por Grecia **e** Irlanda.

Figure 3: Visualization of Confidence Measures

Lisboa y Madrid quieren **emprender** un camino **diferente del adoptado** por Grecia **e** Irlanda.

Figure 4: Interactive Translation Prediction

detecting incorrect words. On the other hand, we highlight in orange color those translated words that are dubious for the system. In this case, we use a threshold that favors recall. See Figure 3 for a screenshot of the text highlighting in the edit area.

Prediction Length Providing the user with a new prediction whenever a key is pressed has been proved to be cognitively demanding (Alabau et al., 2012). Therefore, we decided to limit the number of predicted words that are shown to the user by only predicting up to the first erroneous word according to the CMs.

In our implementation, pressing the Tab key allows the user to ask the system for the next set of predicted words. See Figure 4 for a screenshot.

Search and Replace Most of the computer-assisted translation tools provide the user with intelligent search and replace functions for fast text

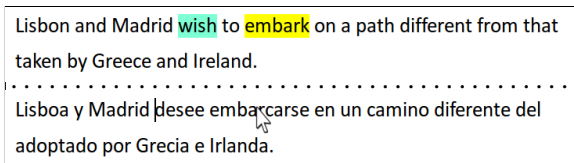


Figure 5: Visualization of Word Alignment

revision. Our workbench features a straightforward function to run search and replacement rules on the fly. Whenever a new replacement rule is created, it is automatically populated to the forthcoming predictions made by the system, so that the user only needs to specify them once.

Word Alignment Information Alignment of source and target words is an important part of the translation process (Brown et al., 1993). In order to display the correspondences between both the source and target words, this feature was implemented in a way that every time the user places the mouse (yellow) or the text cursor (cyan) on a word, the alignments made by the system are highlighted. See Figure 5 for a screenshot.

Prediction Rejection With the purpose of easing user interaction, our workbench also supports a one-click rejection feature (Sanchis-Trilles et al., 2008). This invalidates the current prediction for the sentence that is being translated, and provides the user with an alternate one, in which the first new word is different from the previous one.

2.4 Replay

The workbench implements detailed logging of user activity, which enables both automatic analysis of translator behavior by aggregating statistics and enabling replay of a user session. This capability is explained in detail in Section 4. Replay takes place in the translate view of the UI, it shows the screen at any time exactly the way the user encountered it when he or she interacted with the tool.

2.5 List Documents

Another view details a list of documents submitted to the tool. From there a user can start a replay, download the logged data or continue a translation session.

3 Server

The overall design of the CSMACAT workbench is very modular. There are three independent com-

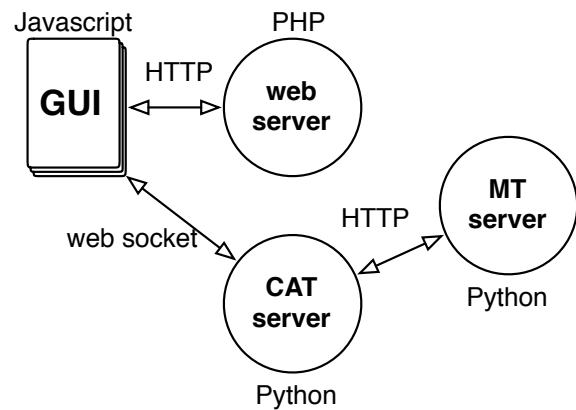


Figure 6: Modular design of the workbench: Web-based components (GUI and web server), CAT server and MT server are independent and can be swapped out

ponents (see also Figure 6): the GUI/web server, the CAT server and the MT server.

We separate these components by clearly specified API calls, so that alternative implementations can be used as well. We expect that the CSMACAT workbench may be use partially, for instance in the following fashion:

- As part of a larger localization workflow with existing editing facilities, only the capabilities of the CSMACAT CAT server and CSMACAT MT server are used. A legacy editing tool is extended to make calls to the CAT server and thus benefit from additional functionality.
- If an existing customized MT translation solution is already in place, then the CSMACAT front-end and CAT server can connect to it.

Already, the currently implemented CSMACAT workbench supports two different MT server components, Moses (Koehn et al., 2007) and Thot (Ortiz-Martínez et al., 2005).

3.1 CAT Server

The CAT server is implemented in Python with the Tornado library. It uses *socket.io* to keep a web socket connection with the Javascript GUI. Keep in mind that especially interactive translation prediction requires very quick responses from the server. Establishing an HTTP connection through an Ajax call every time the user presses a key would cause significant overhead.

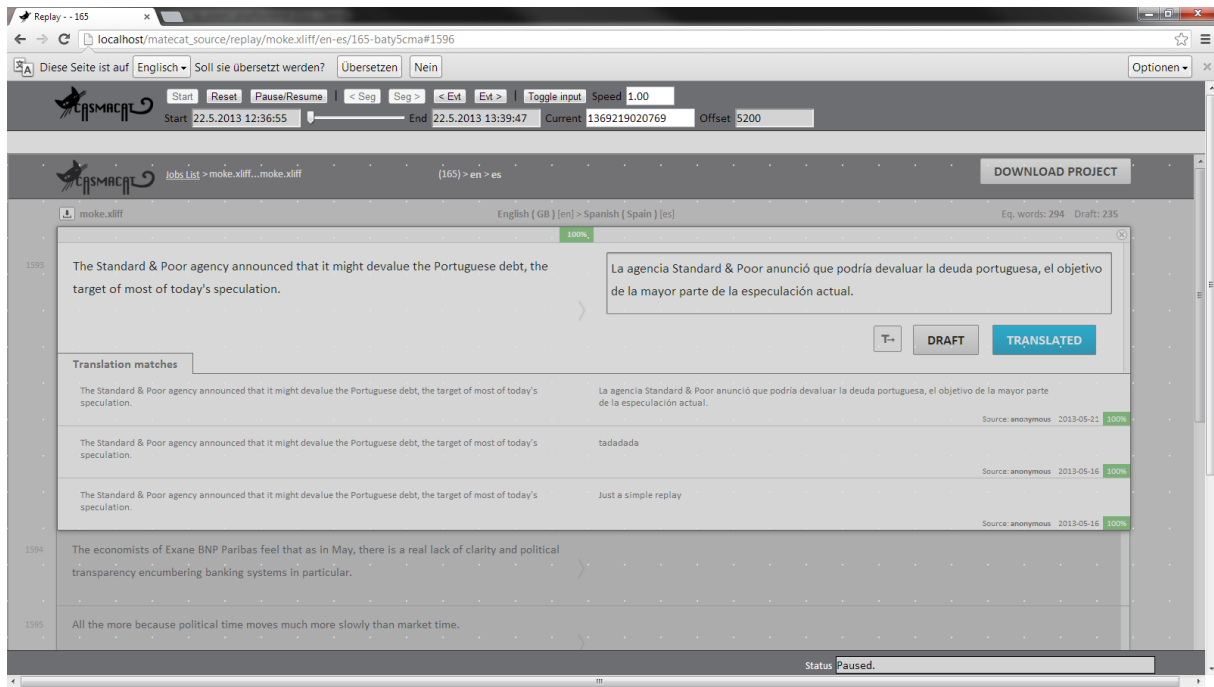


Figure 7: Replay view

A typical session with interactive translation prediction takes place as follows:

- The user moves to a new segment in the GUI.
- The GUI sends a **startSession** request to the CAT tool, passing along the input sentence.
- The GUI and CAT server establish a web socket connection.
- The CAT server requests and receives from the MT server the sentence translation and the search graph.
- The CAT server sends back the translation to the GUI and keeps the search graph in memory.
- The user starts typing (approving some of the translation or making corrections).
- At each key stroke, the GUI sends a request to the CAT server, for instance requesting a new sentence completion prediction (**setPrefix**).
- The CAT server uses the stored search graph to compute a new prediction and passed it back to the GUI (**setPrefixResult**).
- The GUI displays the new prediction to the user.
- Eventually, the user leaves the segment.
- The GUI sends a **endSession** request to the CAT tool.

- The CAT server discards all temporary data structures.
- The GUI and CAT server disconnect the web socket connection.

The interaction between the GUI and the CAT server follows a well-defined API.

3.2 MT Server

For many of the CAT server's functions, information from the Machine Translation (MT) server is required. This includes not only the translation of the input sentence, but also n-best lists, search graphs, word alignments, etc.

The main call to the server is a request for a translation. The request includes the source sentence, source and target language, and optionally a key identifying the user. The server responds to requests with an JSON object, for instance:

```
{
  "data": {
    "translations": [
      {
        "sourceText": "test",
        "translatedText": "testo",
        "tokenization": {
          "src": [[0, 3]],
          "tgt": [[0, 4]]
        }
      }
    ]
  }
}
```

Note that this is based on the API of Google Translate. Our server implementation extends this API in various ways.

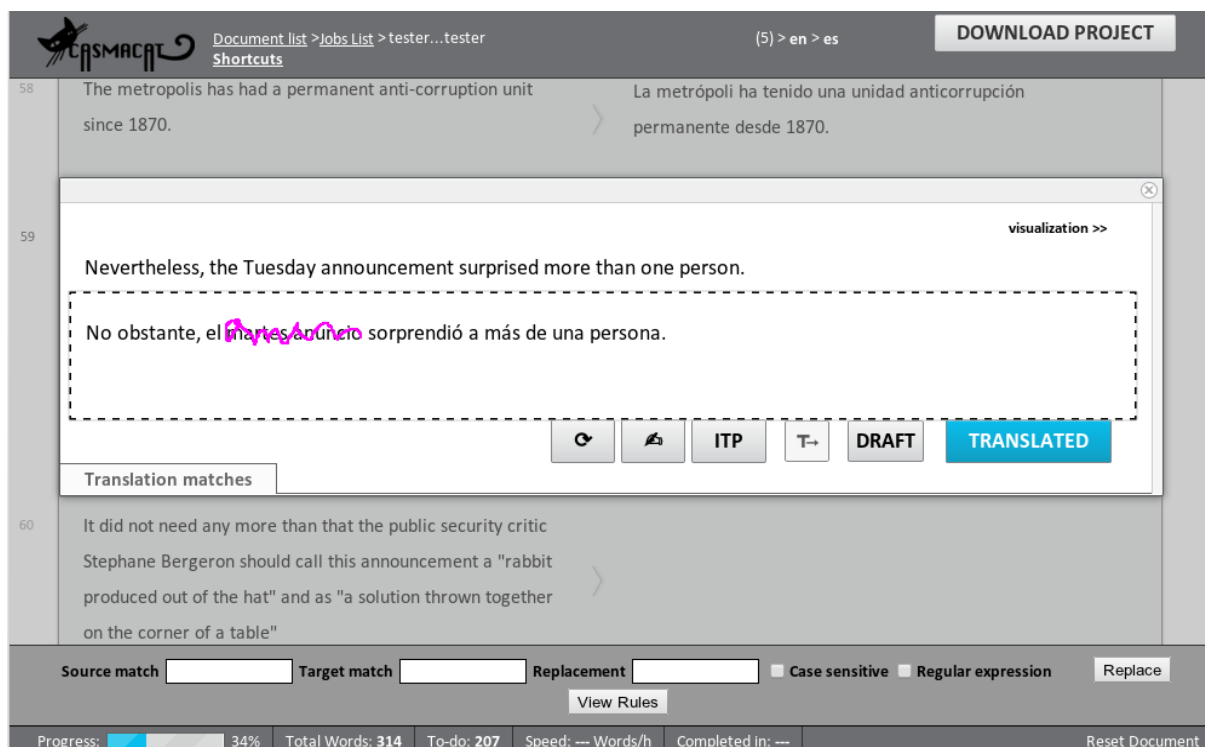


Figure 8: Sketch of a document fragment

4 Replay Mode for User Activity Data

The replay view (see Figure 7) loads the translate view into an *iframe* and remote-controls it with the data from the log file. The session appears in the web browser exactly the same way as it appeared to the user interacting with the tool. The current implementation is robust to user actions, for instance it allows for changes of the replaying window geometry (like resizing).

The log data is fetched in small chunks when replaying. Upon start-up, only the first chunk of the log data is loaded and the replay starts. When the next chunk is needed, the replay is paused and the next chunk is fetched. This minimizes the initial loading time when starting the replay.

The replay engine uses precise internal clocking. Each event is replayed on its on. This makes the engine precise and robust and allows for arbitrary jumps between events.

Many events are included in the logging (such as all events around interactive translation prediction). The functionality is currently being extended to allow for arbitrary seeking in the replay (e.g. by time or segment). Additionally, the replay mode will soon allow to re-compute or re-map par-

ticular data, like gaze-to-char mapping.

Latest tests have confirmed that the current strategy of visualizing the eye tracking data via the browser's DOM is too slow. The new idea is to let the eye tracking plugin take over this task by creating a new native but invisible system window on which the eye tracking data is drawn. This still has to be implemented and tested but promises a high performance visualization of eye tracking data.

5 E-pen Interaction

E-pen interaction should be regarded as a complementary input rather than a complete replacement of the keyboard. In a first approach, we have extended the CSMACAT UI with the minimum components necessary to enable e-pen gestures and handwriting in a comfortable way.

5.1 E-pen UI

When the e-pen UI is enabled, a new button is displayed in the button area (🖋️). This button toggles the e-pen view. When activated, the display of the current segment is changed so that the source segment is shown above the target segment. This way, the drawing area is maximized horizontally, which facilitates handwriting particularly in tablet

devices.

Next, an HTML canvas element is added over the target segment. This drawing area is highlighted with a dashed border. In addition, a clear button (🗑️) is added to refresh the drawing area. A screenshot of such display can be seen in Figure 8.

The user can interact with the system by writing on the canvas. Although in principle it would be interesting to allow the user to introduce arbitrary strings and gestures, in this approach we have decided to focus on usability. We believe that a fast response and a good accuracy are critical for user acceptance.

Thus, we decided to use MINGESTURES (Leiva et al., 2013), a highly accurate, high-performance gestures for interactive text editing. The gestures in MINGESTURES are defined by 8 straight lines that can be configured to be direction dependent and be aware of the context where they gestures takes place. In addition, they can be easily differentiated from handwritten text with line fitting algorithms. Gestures are recognized in the client side so the response is almost immediate.

Conversely, when handwritten text is detected, the pen strokes are sent to the server. At this moment, only single words can be written. However, in future releases also substrings and multiple words will be allowed. The set of gestures used in the workbench are summarized in Figure 9.

5.2 HTR server

The hand-written text recognition (HTR) server is responsible for decoding the user handwriting into digital text. The technology is based very much on the ITP server technology. An HTR server must implement the following API:

startSession This function instructs the server to initialize a new HTR session with the appropriate contextual information. A session consists of one or more strokes that constitute one user interaction. The input parameters are the **source** string, the current **translation** and the last position validated by the user. At this stage, the server does not return a value.

addStroke When a user finishes writing a stroke, the points are encoded into an array of points that are defined by the x and y coordinates along with the *timestamp* when they were acquired. The HTR server processes this infor-

mation and, optionally, returns a partial decoding.

endSession When the user stops writing for a specific amount of time (400ms in our set-up), the users session finishes. The final decoding is then returned to the UI, possibly with a list of n -best solutions.

The HTR server is based on iAtros, an open source HMM decoder. The current version does not leverage contextual information, but it is prepared to support that in future releases.

6 Eye-Tracking

One of the core goals of the CASMACAT project is the study of translator behavior. To better observe the activities of translators, we use eye tracking. This allows us to detect and record the exact screen position of the current focus of attention. Alongside the other logged information such as key logging, enables *translation process study*, i.e., the analysis of the behavior of the translator, opposed to just *translation product study*, i.e., the analysis of the final translation.

Eye tracking is integrated into the CASMACAT workbench using a special plugin for the browser. With this plugin, the eye tracking information is accessible to the Javascript GUI and can be send to the web server to be stored for later analysis. The eye tracking information is also visualized in the replay mode.

Analyzing the eye tracking data requires a tool for aligning and correcting erroneous fixations, to manually map them on the words that are likely to be fixated. The tool is instrumental in creating a corpus of high-quality gaze-to-word mappings, and is used as a training set of automatic gaze-to-word mappings algorithms. A first version of the manual alignment tool was implemented and we also implemented two algorithms for automatic gaze-to-word alignment.

One way to visualize the eye tracking data is by plotting translation sessions in the form of translation progression graphs. However, translation progression graphs only visualize a small fraction of the information.

7 Outlook

We are currently conducting a field trial to extensively test the workbench in a real-world environ-

LABEL	ACTION	RESULT	LABEL	ACTION	RESULT
Substitute		Lorem Ipsan	Split		Lor em
Reject		Lorem	Validate		Lorem Ipsum
Merge		LoremIpsum	Undo		Lorem Ipsum
Delete		Lorem	Redo		Lorem
Insert		Lorem et Ipsum	Help		<help event>

Figure 9: Set of gestures

ment of professional translators. We will collect extensive logging information that will allow analysis of translator behavior and inform the future development of the various technologies.

We hope that by releasing the CASMACAT workbench open source, the broader research community can carry out similar studies.

References

- Alabau, V., Leiva, L. A., Ortiz-Martínez, D., and Casacuberta, F. (2012). User evaluation of interactive machine translation systems. In *Proc. EAMT*, pages 20–23.
- Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., and Vilar, J.-M. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.
- Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.
- Federico, M., Cattelan, A., and Trombetti, M. (2012). Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*.
- González-Rubio, J., Ortiz-Martínez, D., and Casacuberta, F. (2010). On the use of confidence measures within an interactive-predictive machine translation system. In *Proc. EAMT*.
- Koehn, P. (2010). Enabling monolingual translators: post-editing vs. options. In *Proc. NAACL*, pages 537–545.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C. J., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.
- Langlais, P., Foster, G., and Lapalme, G. (2000). TransType: a computer-aided translation typing system. In *NAACL Workshop: EmbedMT*, pages 46–51.
- Leiva, L. A., Alabau, V., and Vidal, E. (2013). Error-proof, high-performance, and context-aware gestures for interactive text edition. In *Proceedings of the 2013 annual conference extended abstracts on Human factors in computing systems (CHI EA)*, pages 1227–1232.
- Ortiz-Martínez, D., García-Varea, I., and Casacuberta, F. (2005). Thot: a toolkit to train phrase-based statistical translation models. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, Phuket, Thailand.
- Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague Bulletin of Mathematical Linguistics*, 93:7–16.
- Pouliquen, B., Mazenc, C., and Iorio, A. (2011). Tapta: A user-driven translation system for patent documents based on domain-aware statistical machine translation. In Forcada, M. L., Depraetere, H., and Vandeghinste, V., editors, *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, pages 5–12.
- Sanchis-Trilles, G., Ortiz-Martínez, D., Civera, J., Casacuberta, F., Vidal, E., and Hoang, H. (2008). Improving interactive machine translation via mouse actions. In *Proc. EMNLP*.
- Skadiņš, R., Puriņš, M., Skadiņa, I., and Vasiljevs, A. (2011). Evaluation of SMT in localization to under-resourced inflected language. In Forcada, M. L., Depraetere, H., and Vandeghinste, V., editors, *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, pages 35–40.

Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE

Joke Daems

Lieve Macken

Sonia Vandepitte

Department of Translation, Interpreting and Communication

Ghent University

Belgium

firstname.lastname@ugent.be

Abstract

Existing translation quality assessment (TQA) metrics have a few major drawbacks: they are often subjective, their scope is limited to the sentence level, and they do not take the translation situation into account. Though suitable for a general assessment, they lack the granularity needed to compare different methods of translation and their respective translation problems. In an attempt to solve these issues, a two-step TQA-approach is presented, based on the dichotomy between adequacy and acceptability. The proposed categorization allows for easy customization and user-defined error weights, which makes it suitable for different types of translation assessment, analysis and comparison. In the first part of the paper, the approach is explained. In the second part of the paper, the approach is tested in a pilot study designed to compare human translation with post-editing for the translation of general texts (newspaper articles). Inter-annotator results are presented for the translation quality assessment task as well as general findings on the productivity and quality differences between post-editing and human translation of student translators.

1 Introduction

With the translation industry exponentially growing, more hope is vested in the use of machine translation (MT) to increase translators' productivity (Rinsche and Portera-Zanotti, 2009). Though

post-editing MT has proven to increase productivity and even quality for certain text types (Tatsumi, 2010), research on the usability of post-editing for more general texts is rather limited. The research presented in this paper is a pilot study conducted as part of the ROBOT-project¹, a project designed to gain insight in the differences between human translation and the post-editing of machine translation. The process and product of translation are the two main areas of interest of the project, and results of student translators and professional translators shall be compared. In this paper, the translation quality assessment approach developed for the project is presented and tested on translations of student translators. This fine-grained, two-step approach does not only allow for the analysis and comparison of translation problems for different methods of translation (such as human translation and post-editing), but can also be used as an evaluation method for different text types and goals. As such, it is a useful tool, both for researchers and people concerned with the evaluation of translation quality in general.

2 Related Research

Although it is the goal of quality assessment schemes and metrics to determine the quality of a translation, the question 'is this a good translation?' can only be adequately answered with the rather vague: 'that depends'. Already more than thirty years ago, Van Slype (1979) established that translation quality is not an absolute concept and thus should be assessed "relatively, applying several distinct criteria illuminating each special aspect of the quality of the translation". Though the

¹lt3.hogent.be/en/projects/robot

focus of his report was on the evaluation of machine translation, the definition holds true for every type of translation.

Since then, a lot of translation quality assessment schemes have been proposed, most of them based on an error typology. Some examples include the SAE J2450 (2001), LISA (2011), and EN-15038 (2006). Though useful in certain contexts, these typologies have three major drawbacks that limit their usability for the ROBOT-project. First of all, they are usually designed for a specific text type or domain and can not easily be tailored to different text types. The J2450, for example, is used in the automotive sector and has a limited amount of categories. Secondly, they do allow for the integration of severity scores, but these are often subjective. There is a distinction between ‘minor’ and ‘major’ errors, and sometimes a third ‘critical’ category is added, but no real rules are given on how to discern between a ‘minor’ and a ‘major’ error. And finally, the categorization is not fine-grained enough to allow for a thorough analysis between different methods of translation.

In order to create a more generally applicable translation quality assessment scheme, it seems wise to look at a general definition of what constitutes a good translation. According to Chesterman (1998), prototypical translation consists of the following features (among others): the intended function, text type, and style of TT are similar to those of the ST; the TT renders all contents of the ST; and the style of the TT is ‘good native’. This type of translation requires a high level of fidelity towards the source text while at the same time being fluent and grammatically correct in the target language. Adherence to the norms of the source text while at the same time respecting the norms of the target text are what Toury (1995) calls adequacy and acceptability, respectively. This distinction is often used in assessment schemes for MT-quality (White, 1995). One of the problems with these schemes, however, is that they are usually restricted to the sentence-level and don’t take general coherence problems into account. And, just as with the previously mentioned quality metrics, the severity scores are often subjective.

More recently, researchers have tried to overcome the flaws of previous metrics by paying more attention to the goal of the source text and target text. Williams (2009) suggests using an argu-

mentation centred translation quality assessment to make sure the macro structure of the source text is respected, whereas O’Brien (2012) opts for a dynamic approach to quality evaluation, taking into account the goal of the translation, the time and the resources of the company. Colina (2009) introduces an assessment tool with a functionalist approach, allowing for a user-defined notion of quality. Though the idea behind the tool is in line with the need for a more flexible approach to translation quality assessment, the tool’s usefulness is limited to providing a quick assessment of the main problem categories of a text, but it lacks an in-depth analysis of the types of errors and gives no concrete suggestions for improvement.

The translation quality assessment approach presented in this paper was designed to overcome many of the problems encountered when using existing translation quality assessment metrics. An overview of the approach is given in the following paragraphs, followed by a pilot study during which the approach was tested.

3 A two-step TQA approach

Starting from Chesterman’s definition of prototypical translation (1998) as the baseline goal of a translation, an evaluation categorization was designed, consisting of acceptability and adequacy as main categories. The idea is that a translation requester (be it a teacher or a company) would want a translation to be both a fluent, correct text in the target language, as well as a text that conveys all the information contained in the source text in an appropriate way. Rather than just giving a generic ‘acceptable’ vs. ‘unacceptable’ assessment for both categories, the categories are further subdivided in order to be able to discern specific translation problems. This approach allows teachers to provide in-depth feedback to their students, researchers to analyse differences between text types or translation methods (MT+PE vs. HT, for example), and clients to easily revise a translation.

In a preliminary test of the approach, revisers had to annotate problems for adequacy and acceptability at the same time. This strategy, however, had a few drawbacks. There was some confusion on the appropriate category for the problem at hand (was it caused by adequacy problems or was it acceptability-related?), and annotators lost track of

the coherence of the text because source and target sentences were alternated. The solution to these problems lay in dividing the process into two steps. In a first phase, annotators get to see the target text without the source text and they have to annotate the text for acceptability only. In a second phase, they get to see the source sentences alternated with their translations and they have to annotate these sentences for adequacy only.

3.1 Categorization

For acceptability, the main categories consist of grammar and syntax, lexicon, spelling and typos, style and register, and coherence. While the first three categories build on existing TQA metrics, the second two are less common. They have been included in order to be able to identify problems related to the text in context and the text as a whole. The subcategory ‘text type’ is used to highlight genre-specific problems such as the use of articles in newspaper titles. Previous metrics often did not take the goal of the text into account or were meant to assess texts sentence per sentence, thus losing the overview and coherence. The subcategories of each category can be found in Table 2 below. The numbers between brackets indicate the proposed error weight for the translation of general texts, which will be further explained in the following paragraphs. For adequacy, the main category, viz. meaning shift, is further subdivided in different subcategories (see Table 1).

Meaning shift
contradiction (3)
word sense disambiguation (3)
hyponymy (1)
hyperonymy (1)
terminology (0)
quantity (2)
time (2)
meaning shift caused by punctuation (2)
meaning shift caused by misplaced word (3)
deletion (2)
addition (2)
explicitation (0)
coherence (2)
inconsistent terminology (0)
other (2)

Table 1: Adequacy subcategories with error scores

These categories may seem rather fine-grained, but this allows for a more thorough analysis of translations, not just for quality assessment. Deletions, for example, are expected to be more common in human translations than in post-editing, but it could be interesting to see when one opts for a hyponym or hyperonym rather than a straightforward translation as well. In the same fashion ‘explicitations’ are usually not considered to be errors, but they do provide interesting information on the translation process. A ‘meaning shift caused by misplaced word’, on the other hand, would be considered to be an error. This occurs when the words are correctly translated, but they are connected in a wrong way. For example, when a translator interprets a sentence about ‘unorthodox cancer cures’, as being about ‘unorthodox cancer’ rather than about ‘unorthodox cures for cancer’. A more detailed overview of all categories with examples can be found in (Daems and Macken, 2013).

It must be noted at this point that the proposed categorization does not claim to be exhaustive. It was primarily designed for use within the ROBOT-project, for the translation of English texts into Dutch, with the main focus on the translation of general texts. This, however, does not mean that the categorization has a limited use. By using the well-known distinction between adequacy and acceptability and universal concepts such as ‘grammar’ and ‘lexicon’, this categorization can easily be tailored to suit language-specific problems.

Important to keep in mind as well, is the fact that this categorization provides an overview of *possible* translation problems. While grammatical problems will most likely be considered to be errors in most cases, the line for other categories is less clear. Depending on the goal of the text and the goal of the evaluation, certain problems will or won’t be regarded as an error, but rather as translation characteristics. This principle will be further explained in the following paragraphs.

3.2 Objectivity & Flexibility

As became clear from the related research, translation quality assessment approaches should on the one hand be more dynamic in that they should take the translation situation and context into account and on the other hand be more objective in their value judgement. The proposed TQA approach tries to fulfil these requirements by allowing for

Grammar & Syntax	Lexicon	Spelling & Typos	Style & Register	Coherence
article (2)	wrong preposition (2)	capitalization (1)	register (1)	conjunction (3)
comparative/superlative (2)	wrong collocation (2)	spelling mistake (1)	untranslated (2)	missing info (3)
singular/plural (2)	word nonexistent (2)	compound (1)	repetition (1)	logical problem (3)
verb form (2)		punctuation (0)	disfluent (1)	paragraph (2)
article-noun agreement (2)		typo (0)	short sentences (1)	inconsistency (2)
noun-adj agreement (2)			long sentence (1)	coherence - other (3)
subject-verb agreement (2)			text type (2)	
reference (2)			style – other (2)	
missing (2)				
word order (2)				
structure (2)				
grammar – other (2)				

Table 2: Acceptability subcategories with error scores

user-defined categorizations and error weights.

Depending on the goal of the translation or the evaluation, the user adopts the proposed categorization as is or adapts it to better suit his wishes and/or language pair. The most important aspect of the evaluation, however, is the addition of error weights. Unlike with existing TQA schemes, the error weights for the current approach are not predefined or intuitively added by the reviser. It is the user who decides on the error weight for each subcategory. The error weights used for the current paper have been adapted to the translation of newspaper articles from English to Dutch, and can be found in Table 1 for adequacy and in Table 2 for acceptability. The main idea is that problems that have a larger impact on readability and comprehension receive a higher error weight. Depending on the goal of the assessment, the user can decide to change the error weights as desired. In technical texts, for example, the category ‘terminology’ would receive a high error weight. It is even possible to give no weight to a category. This is especially useful to detect differences in translations, without these differences necessarily being errors, such as explicitation or hyperonyms, which in turn could be interesting to examine differences between - for example - human translation and post-editing.

As the translation environment we used for the experiment did not contain a spell-check function, the subcategories ‘punctuation’ and ‘typos’ were also assigned a zero weight.

4 Experiment

To test the proposed categorisation and two-step TQA approach, a pilot experiment was conducted in which participants had to both translate a text and post-edit a machine-translated text from English to Dutch. The goal of the experiment with regards to the proposed TQA approach was twofold: to check whether or not the guidelines were sufficiently detailed for the annotators, and to check whether the approach is a viable tool for a comparative analysis of translation problems for different methods of translation.

4.1 Experimental set-up

Participants were 16 Master’s students of translation taking a general translation course. Students had no experience with post-editing and were given no specific training. Each student received a translation and a post-editing task. The machine translation to be post-edited was obtained by using Google Translate². The order in which the students received the tasks differed, to reduce task order effects. There was no time restriction. The corpus consisted of four newspaper articles of more or less equal length (260-288 words) taken from the Dutch Parallel Corpus (Macken et al., 2011). The instruction for both tasks was to achieve a translation of publishable quality, and the target audience of the translations was said to be more or less equal to the target audience of the source text. Participants were informed that they would receive feedback on the productivity and quality of their translations.

²translate.google.com

The text difficulty was estimated by uploading the texts on editcentral.com, which provided scores for six different readability indexes. According to these scores, the first two texts (1722 & 1771) were slightly less difficult than the last two (1781 & 1802), with Flesch-Kincaid levels of 10.7 and 12.4, and 16.5 and 14.6 respectively.

The tasks were recorded with PET, a post-editing tool developed by Aziz and Specia (2012), which allows for keystroke logging and time registration. The original English text was presented on the left hand side of the screen, whereas the right hand side was empty for the regular translation task, or showed the Dutch MT output for the post-editing task. Only one sentence at a time could be edited, but the four previous and next segments were always visible so that students could take the context into account. They were also allowed to go back to revise segments they had already translated or post-edited. Each sentence was followed by an assessment screen in which the students commented on the external resources they consulted and assigned a subjective difficulty score to each sentence.

4.2 Translation Quality Assessment

All translation and post-editing products were annotated by two annotators, according to the guidelines and the categorization introduced above. The annotators were translation and language specialists (one with a Master's degree in Translation (English-Dutch) and one with a Master's degree in English and Dutch linguistics). For the annotations, the brat rapid annotation tool was used (Stenetorp et al., 2012). In this tool, users can add their own annotation scheme and texts. It provides a nice interface and user-friendly environment, so no real annotator training was needed. The annotators had to follow the guidelines published in (Daems and Macken, 2013). In a first phase, annotators had to annotate all products for acceptability (in which case they only received the target text). In a second phase, annotators had to annotate the products for adequacy (in which case they received a text where source and corresponding target sentences were alternated). The annotation tasks were presented in a random order, with at least two different texts between every two products from the same source text.

Before starting with the annotations, annotators

were informed about the translation task and purpose. They were instructed to highlight those items that were (either linguistically or conceptually) incorrect with regards to the text type and the audience and to provide a short comment on the reason for each annotation. In case of doubt, annotators were asked to add a double question mark to their comments. This facilitated the automatic analysis of the final data.

4.3 Inter-annotator agreement

To determine the validity of the approach, two aspects had to be examined: Do annotators highlight the same items? And if they do, do they label the items with the same category? This was tested by calculating inter-annotator agreement over all texts for both acceptability and adequacy in different stages. The initial agreement was calculated on the basis of the annotations as initially received from both annotators. Of the 796 acceptability annotations, only 341 were highlighted by both annotators. This led to an agreement of 38% with $\kappa=0.31$. For adequacy, only 134 of the 291 cases were highlighted by both annotators, equal to an agreement of 41% with $\kappa=0.30$. Though these numbers are rather low at first sight, a few things must be taken into account. First of all, certain errors were highlighted by only one annotator simply because the other annotator hadn't observed the error, not because the annotator did not agree with the judgement. Secondly, some errors recurred in different translations, so a disagreement on one conceptual item could lead to a large difference when looking at the number of annotations. A linear regression was fitted to verify whether or not the annotators' overall assessment was the same, as it was hypothesized that a 'strict' annotator would be equally strict across all texts, and a more 'lenient' annotator would be equally lenient across all texts. This hypothesis was confirmed as we found a positive correlation for both adequacy and acceptability annotations, $r=0.89$, $n=38$, $p<0.001$ and $r=0.70$, $n=38$, $p<0.001$, respectively. Moreover, when looking at the items that were highlighted by both annotators, it seems that agreement on the categories is rather high: 89% with $\kappa=0.88$ for acceptability and 89% with $\kappa=0.87$ for adequacy, which indicates that the categorisation itself seems to be rather clear.

A consolidation phase was introduced to check

whether or not annotators agreed with each other's annotations. This phase was twofold: firstly, a manual phase was introduced to extract those cases where the annotators identified the same problems, but labelled them differently, and secondly, a list was made of all the annotations labelled by only one annotator. In consultation with the annotators, a final category was assigned to each of the problems that had received different labels. Most of these cases were caused by ambiguity in the guidelines or non-adherence to the guidelines by one of the annotators. Where possible, the guidelines were further disambiguated and more examples were added to overcome these problems in the future. For the second step, the annotators received a list of all the annotations that were only labelled by the other annotator. They had to indicate whether or not they agreed with the annotations. Agreement after the consolidation phase was much higher: 69% with $\kappa=0.67$ for acceptability and 82% with $\kappa=0.79$ for adequacy.

This final set of annotations after the consolidation phase forms the gold standard and is used in the next section to analyse the differences between post-editing and human translation.

4.4 Results

The goal of the pilot study was to analyse differences between the post-editing of machine translation and human translation for the translation of newspaper articles by student translators. More specifically, it was concerned with differences in productivity as well as quality.

4.4.1 Productivity

The productivity for each text and each type of translation was measured by the PET post-editing tool. As can be seen in Figure 1 below, post-editing was always faster than human translation.

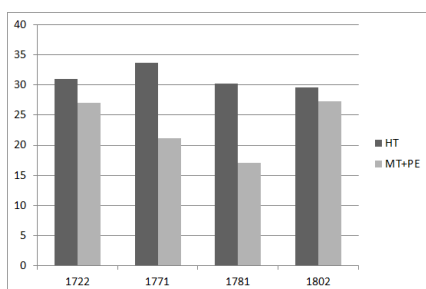


Figure 1: Time spent per text in minutes

These results seem to support the findings of Guerberof (2009) and Plitt and Masselot (2010) that post-editing machine translation can lead to an increase in productivity, compared to regular translation. Of course, an increase in productivity is only positive when the quality does not suffer from the higher speed.

4.4.2 Quality: totals

The average error score for acceptability and adequacy for each type of translation per text can be found in Figures 2 and 3 below. The score is calculated by taking the sum of all annotated problems in the gold standard (annotations after consolidation phase) multiplied by their respective error weights (see Tables 1 and 2).

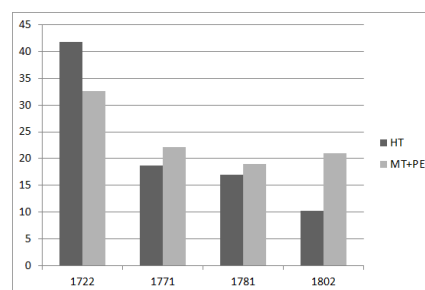


Figure 2: Average acceptability error score per text

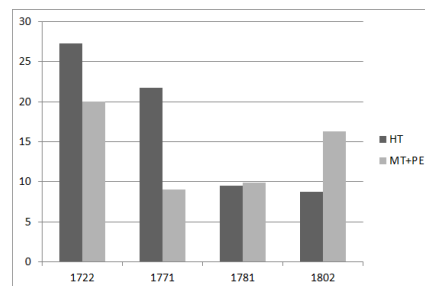


Figure 3: Average adequacy error score per text

What can be derived from these graphs is that quality is extremely text-dependent. For text 1722, acceptability quality is much higher for post-editing than for human translation, whereas the opposite can be said of text 1802. For texts 1771 and 1781 the acceptability quality is slightly higher for human translation than for post-editing. When looking at adequacy, it can be seen that quality is much higher for texts 1722 and 1771 for post-editing in comparison with human translation. The difference for text 1781 is negligible, but for text

1802, human translation seems to lead to higher adequacy than post-editing.

To calculate the total error score for each text, it was not possible to simply add up the adequacy and acceptability scores, because quite a few problems were annotated both as acceptability and as adequacy problems. In these cases, acceptability problems resulted from a mistranslation or other adequacy issue, so it was decided that only the error weight for the adequacy annotation would count. The average of the total error scores thus obtained can be seen in Figure 4.

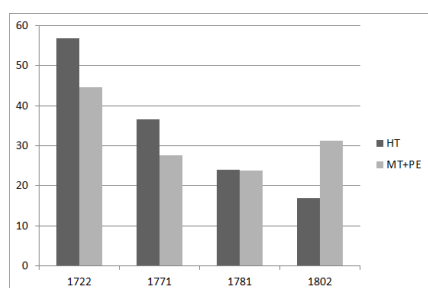


Figure 4: Average total error score per text

Overall, it seems that post-editing often leads to higher quality than human translation. This is true for three of the four texts, though the difference for text 1781 is once again negligible. No significant correlation was found between translation or post-editing time and error score.

4.4.3 Quality: problem analysis

Though the totals for adequacy and acceptability already provided some insights in the differences between post-editing and human translation, the main goal of the proposed categorisation was to allow for a more thorough analysis of translation problems. A more detailed picture is given by the analysis of the main subcategories. As was the case for the difference between adequacy and acceptability, the scores for each category depend largely on the text. When looking at the most common problems for each text, it becomes clear that ‘meaning shift - other’ and ‘meaning shift - deletion’ are very common categories for human translation for texts 1722 and 1771 (with ‘other’ taking up 9% and 15% of total human translation error for these texts and ‘deletion’ accounting for 7% and 15% of total human translation errors), but not so common for post-editing (the categories ‘other’ and ‘deletion’ were not found in text 1722 and only

accounted for 3% and 5% of all post-editing errors for text 1771, respectively). Common post-editing problems, on the other hand, seem to be ‘meaning shift - wrong word sense’ and ‘lexicon - wrong collocation’. ‘Wrong word sense’ accounted for 14% of all PE-errors for text 1722 (for HT, this was a mere 5%), 9% of all PE-errors for text 1781 (versus 1% for HT), and 10% of all PE-errors for text 1802 (compared to 4% for HT). ‘Wrong collocation’ accounted for 17% of all PE-errors for text 1722 (compared to 7% for HT), 9% of all PE-errors for text 1771 (compared to 5% for HT) and 17% for text 1781 (compared to 7% for HT). Some categories were only important issues for one of the four texts, such as ‘compounds’ for text 1722 (with a HT and PE value of 4% and 8% respectively), ‘capitalization’ for text 1771 (with a HT and PE value of 8% and 4% respectively) and ‘register’ for text 1802 (with a HT and PE value of 7% and 1% respectively). A more thorough analysis of these differences could yield insights in text differences and what makes a text difficult to translate (both for a human and a machine), but this would go far beyond the scope of the present article.

When looking at the global overview of the three most common categories for human translation and post-editing, depicted in Figure 5 below, it can be derived that especially meaning shifts are common problems for human translation, whereas post-editing suffers most from wrong word sense disambiguation and wrong collocations. Though it is common for a machine translation system to select the wrong meaning of a word, it is remarkable that these errors are not spotted by the post-editors.

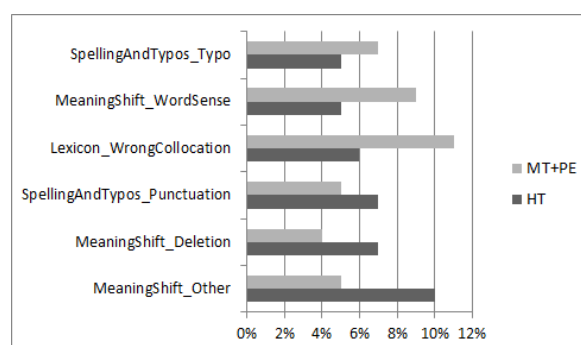


Figure 5: Most common error types over all texts

Figures 6 and 7 provide an overview of the proportion of each problem category in relation to the total amount of problems, both for HT and

MT+PE. In Figure 6, the focus is on the sub-categories of acceptability and adequacy is represented as one large sector. In Figure 7, on the other hand, the most important adequacy categories are highlighted and acceptability is represented as one sector. The categories that did not show remarkable differences have been grouped in ‘Adequacy_grouped’.

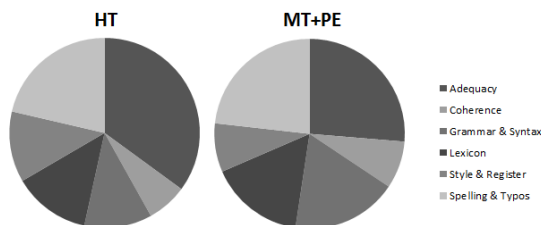


Figure 6: Proportion of problem categories: focus on acceptability

The largest proportion of problems is accounted for by acceptability for both types of translation. Whereas the bulk of acceptability errors seems to be caused by spelling errors, there are some differences between HT and MT+PE: A large proportion of post-editing problems is caused by grammar & syntax and lexical problems, while for human translation style & register issues seem to be more common.

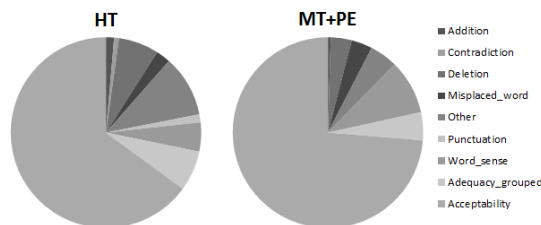


Figure 7: Proportion of problem categories: focus on adequacy

Adequacy as a whole accounts for a large percentage of total HT-problems, whereas this amount is noticeably lower for MT+PE. Remarkable as well is the fact that there are more different types of adequacy errors for human translation than for post-editing: contradiction and punctuation problems were only found in human translations. Other than this, it can be derived from Figure 7 that additions and deletions are more common for human translations, along with ‘other’ types of meaning shifts, which take up a large portion of the total

amount of adequacy errors for human translation. Categories that are clearly more common for post-editing are ‘word sense’ and ‘misplaced word’.

5 Discussion & Future work

In this paper, a new, two-step TQA-approach was presented, designed for a detailed analysis of translation problems. The approach is based on the distinction between adequacy and acceptability and the error classification and user-defined error weights allow for adaptation to different text types and assessment goals. The usability of the approach was validated in a pilot study with master’s students of translation, where it was used to on the one hand define the quality of translations and on the other hand provide a deeper understanding of the differences between human translation and post-editing of general texts. Seeing as the experiment was a pilot study, only cautious conclusions can be drawn, yet the study led to some important findings and interesting directions for future research. Firstly, there is a large amount of annotations made by only one annotator, which highlights the need for more than one annotator when assessing translation quality and the need for clear guidelines and briefing. Secondly, quality is highly text-dependent, so different texts should be analysed before conclusions can be drawn. Thirdly, post-editing is faster than human translation, while at the same time being of comparable quality, depending on the text. It is hypothesised that by training post-editors to detect typical PE-issues (such as word sense, grammatical problems and wrong collocations) the quality of post-editing can be increased still. An important remark is the fact that the pilot study was conducted with translation students of different levels (although they were all Master’s students from the same year), and experiments with professional translators could lead to different results. Furthermore, the annotation process is a rather time-consuming process, so the need of more than two evaluators should be carefully considered, depending on the requirements of the project. Within the framework of the ROBOT-project, two annotators proved to be sufficient in that their annotations after consensus allowed for an in-depth analysis and comparison of HT and PE texts. Other plans for future work include comparing the proposed TQA-approach to different methods of TQA, linking the differences in translation

quality to the original MT-quality (in order to better understand post-editing problems), and linking the differences in translation quality to text difficulty (as readability scores do not seem to indicate translatability, so more research in this field is required as well). A final goal for future research is the application of the proposed TQA-approach to different text types and perhaps languages, to prove its adaptability to different situations.

References

- Aziz, Wilker, Sheila de Sousa, and Lucia Specia. 2012. PET: a tool for post-editing and assessing machine translation. *LREC 2012, The 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey. May 2012.
- Chesterman, Andrew. 1998. *Causes, Translations, Effects*, Target, 10(2):201-230.
- Colina, Sonia. 2009. *Further Evidence for a Functional Approach to Translation Quality Evaluation*, Target 21(2):235-264.
- Daems, Joke, and Lieve Macken. 2013. *Annotation Guidelines for English-Dutch Translation Quality Assessment, version 1.0*. LT3 Technical Report - LT3 13.02. available from lt3.hogent.be/en/publications/annotation-guidelines-for-english-dutch-translation-quality/
- EN 15038. 2006. *Translation services - Service requirements*
- Guerberof, Ana. 2009. Productivity and quality in MT post-editing. *MT Summit XII - Workshop: Beyond Translation Memories: New Tools for Translators MT*, Ottawa, Ontario, Canada.
- Localization Industry Standards Association. LISA QA Model 3.1. available from www.lisa.org/LISA-QA-Model-3-1.124.0.html
- Macken, Lieve, Orphée De Clercq, and Hans Paulussen. 2011. *Dutch Parallel Corpus: a Balanced Copyright-Cleared Parallel Corpus*, Meta, 56(2): 374-390. Les Presses de l'Université de Montréal.
- O'Brien, Sharon. 2012. *Towards a Dynamic Quality Evaluation Model for Translation*, The Journal of Specialised Translation(17):55-77.
- Plitt, Mirko and François Masselot. 2010. *Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context.*, The Prague Bulletin of Mathematical Linguistics, 93:7-16.
- Rinsche, Adriane and Nadia Portera-Zanotti. 2009. *The size of the language industry in the EU*. Retrieved from http://ec.europa.eu/dgs/translation/publications/studies/index_en.htm
- SAE J2540. December 2001. *Quality Metric for Language Translation*. www.apex-translations.com/documents/sae_j2450.pdf
- Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. *EACL 2012, Proceedings of the Demonstrations Session at the 13th European Chapter of the Association for Computational Linguistics.*, Avignon, France.
- Tatsumi, Midori. 2010. *Post-Editing Machine Translated Text in a Commercial Setting: Observation and Statistical Analysis*. Dublin: Dublin City University.
- Toury, Gideon. 1995. *The Nature and Role of Norms in Translation*, Descriptive Translation Studies and Beyond:53-69. Amsterdam-Philadelphia: John Benjamins.
- Van Slype, Georges. 1979. *Critical Methods for Evaluating the Quality of Machine Translation*. Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management, Report BR-19142. Bureau Marcel van Dijk.
- White, John. 1995. Approaches to Black-box Machine Translation Evaluation. *Proceedings of the MT Summit 1995*, Luxembourg.
- Williams, Malcolm. 2009. *Translation Quality Assessment*, Mutatis Mutandis 2(1):3-23.

Transferring Markup Tags in Statistical Machine Translation: A Two-Stream Approach

Eric Joanis, Darlene Stewart, Samuel Larkin and Roland Kuhn

National Research Council Canada

1200 Montreal Road, Ottawa ON, Canada, K1A 0R6

First.Last@nrc-cnrc.gc.ca

Abstract

Translation agencies are introducing statistical machine translation (SMT) into the work flow of human translators. Typically, SMT produces a first-draft translation, which is then post-edited by a person. SMT has met much resistance from translators, partly because of professional conservatism, but partly because the SMT community has often neglected some practical aspects of translation. Our paper discusses one of these: transferring formatting tags such as **bold** or *italic* from the source to the target document with a low error rate, thus freeing the post-editor from having to reformat SMT-generated text. In our “two-stream” approach, tags are stripped from the input to the decoder, then reinserted into the resulting target-language text. Tag transfer has been tackled by other SMT teams, but only a few have published descriptions of their work. This paper contributes to understanding tag transfer by explaining our approach in detail.

1 Introduction

Increasingly, translation agencies are incorporating machine translation (MT) in the work flow of their translators, to make them more productive. The usual scenario is a variant of post-editing, in which an initial translation generated by MT is manually corrected by a human translator (who is now called a “post-editor” instead). Green et al. (2013) show that SMT followed by post-editing can improve translator productivity, and even translation quality. Some interesting questions arise. For instance, how should MT interact with other productivity tools used by translators, such

as terminology databases and translation memories? Koehn and Senellart (2010) and Du et al. (2010) discuss these issues.

Many translators resist using MT. Green et al. (2013) write bluntly: “Translators often show intense dislike for working with MT output.” This is confirmed by our own experience. Why does this particular productivity tool, unlike others mentioned above, attract so much hostility? Perhaps it is because of status anxiety: a translator who becomes a post-editor may perceive him/herself as having been proletarianized, going from being the machine’s master to its slave.

However, there are also practical objections to putting out-of-the-box MT into the translation workflow. This paper looks at one of these, the need to transfer markup tags from the source to the target document. Tags are omnipresent in real-life documents: they appear in almost every file format (e.g., Word, HTML, Excel, PowerPoint, PDF) and are used to encode everything that is not plain text (fonts, footnotes, table cells, hyperlinks, etc.). In the translation memories of our clients, 10–40% of segments have at least one tag. In sentences having tags, the average number of tags was about 3 per sentence.

In this paper, we will only discuss tag transfer for language pairs with similar orthographies, especially European languages (e.g., it’s not clear what the equivalent of “italic” would be in Chinese text). We will only discuss tag transfer for statistical MT (SMT) systems: rule-based MT systems became translators’ tools decades ago, and their designers worked out then how to handle this problem. We are unfamiliar with their solutions (finding out details about commercial rule-based systems tends to be difficult). By contrast, SMT researchers often give themselves the luxury of pretending that only pure text matters (but see “Related Work”). Unfortunately, this is not the world in which professional translators

live. For them, post-editing SMT output without the formatting information found in the source may represent a serious loss of productivity.

There are two possible approaches to the problem. One might consider formatting to be an intrinsic part of the source text, carrying important information that should help determine the choice of words and word order in the translation. In that case, tags should be part of the information available to the SMT decoder. This is the “one-stream” approach. By contrast, in the pure “two-stream” approach we adopted, all text fed to the decoder has been stripped of formatting information; once a target-language translation has been produced, the tags are reinserted into it. **Section 3** below discusses how previous work by other authors fits into this classification. An original aspect of our implementation is that the tag re-insertion module has information not only about phrase pair alignments, but also about word alignments within each phrase pair.

There are tradeoffs between the one- and two-stream approaches. The two-stream approach pools information in the SMT system’s training data more efficiently: e.g., the word sequences “he never wins”, “he **never** wins”, and “he never wins” will look exactly the same in the training data, instead of differing because of the tags around the central word. (Data pooling could be achieved in the one-stream approach too, but at the cost of complicating decoding). An advantage and a disadvantage of the two-stream approach is that it permits the decoder to break apart words found together inside a paired tag: the decoder has more freedom to choose a good translation, but that may make deciding where to put the tag in the target text more difficult.

In the two-stream approach, the decoder may initially translate “**Hang up** the phone!” into German as “Lege das Telefon auf!” Depending on how the tag reinsertion rules are written, the final translation might be “Lege das Telefon auf!”, “**Lege das Telefon auf!**”, or even “Lege das Telefon auf!” In the one-stream approach, we can easily tell the decoder not to break up a contiguously-tagged word sequence. The decoder would probably produce “Lege auf das Telefon!”, which exactly reproduces the formatting of the original, but has unidiomatic word order.

This paper gives a detailed description of our implementation of the pure two-stream approach, to clarify the issues and to help other people who might wish to implement it. We do not carry out

an experimental comparison between the one-stream and two-stream approaches, though this would certainly be a worthwhile next step.

2 Background

2.1 Tags and Translator Productivity

Consider translating this sentence into French:

Acknowledgement section should go as a last section immediately *before the references*.

An old-fashioned translator without access to a translation memory might “translate by replacement”: put French text into a copy of the source, progressively deleting English words. This transfers the format, because word processing programs typically transfer formatting to adjacent characters. The translator might begin by typing the French word “Remerciements” next to its English equivalent, “**Acknowledgement**”, so the text may briefly look like this:

AcknowledgementRemerciements section should go as a last section immediately *before the references*.

Next, the translator will delete “**Acknowledgement**” and proceed to type other French words in place. Virtually no productivity is lost by coping with the **bold** font of “**Acknowledgement**” (nor, later on, with the *italics* of “*before the references*”).

Often, users of translation memories get format transfer for free: if words they keep in the target sentence retrieved from the memory are formatted, perhaps that formatting is still appropriate. This is not guaranteed to happen, but can occur in cases where translations are regularly updated (e.g., with Web pages).

If the translator uses a translation provided by vanilla SMT, neither of these tag transfer shortcuts is possible. For example, suppose the source sentence above yields the following translation (taken from vanilla Google Translate):

Section de reconnaissance doit aller une dernière section immédiatement avant les références.

Here, not only must the post-editor fix up the text of the translation, but he/she may also need to change the typeface for the whole French sentence, put part of it into **bold**, and put another part into *italic*. Each manipulation represents a further loss of productivity.

In the example just given, we would like the tag transfer module to modify the output into something like this (before post-editing):

Section de **reconnaissance** doit aller une dernière section immédiatement *avant les références*.

2.2 Markup Formats and Wrappers

In SMT applications for translation agencies, both the source sentence and the training data for the SMT system are typically in markup formats used by translation memories. Two common open-standard formats for these are Translation Memory eXchange (TMX) (LISA, 2005) and XML Localization Interchange File Format (XLIFF) (OASIS, 2008), XML standards for storing and transferring translation memory contents, and documents to be or already translated, between applications.

A translation work package is a unit of work a manager would give a translator—typically, all sentences that need translating in a document from the client, but sometimes part of a document or several documents—and the matching sentences in the other language (where available). XLIFF was designed to transfer a document with its translations (localized versions), while TMX was designed to transfer whole translation memory contents (which could be thousands of documents from hundreds of translators). Though both formats can be used to store a translation work package, XLIFF is better suited for that purpose because it was designed with that end in mind. It is establishing itself as the standard for translation work packages, both in proprietary software (e.g., the SDL Trados suite) and in open source translation memory software (e.g., the Okapi Framework).

In principle, the markup format is irrelevant to the work described in this paper. We used files exported by SDL Trados, the dominant commercial provider of translation memories. When Trados exports TMX and XLIFF files (or files in any one of over 70 formats), it hides the markup details in a wrapper layer (SDL International, 2011). Since TMX tags are still complex, we re-wrap them in another simplifying layer. From our system’s perspective, there are two types of tags: isolated tags and paired tags—it has no notion of italic, bold, typeface, etc. Isolated tags occur when the underlying document has a point tag, or when it has a tag pair that starts in one

sentence and ends in another: in that case we see an isolated tag in each sentence.

In practice, though our system handles TMX and XLIFF files in the same way, the formatting of a given sentence is often more economical in XLIFF than in TMX. For instance, in our XLIFF files, tags that are common to a sequence of words tend to be factored out more than in TMX files.¹ Thus, our system often generates better output from XLIFF than from TMX, even though both are in wrapped formats, because it gets confused by the TMX’s verbosity.

Below is a real example of TMX and XLIFF tags for the same text. The id numbers after the “=” point to the formatting details (e.g., **bold**, *italic*, etc.):

Original text: National CH₄ and N₂O emissions.

TMX: <op id="6"/><op id="7"/>National CH
<cl id="7"/><cl id="6"/><op id="8"/>4
<cl id="8"/><op id="9"/> <op id="10"/>and
N<cl id="10"/><cl id="9"/><op id="11"/>2
<cl id="11"/><op id="12"/>O emissions.
<cl id="12"/>

NOTE: <op...> and <cl...> stand for our wrappers around the actual “open” (<bpt>...</bpt>) and “close” (<ept>...</ept>) tags in the TMX.

XLIFF: National CH<g id="3173">4</g> and
N<g id="3174">2</g>O emissions.

NOTE: Here <g id=...> and </g> are the “open” and “close” tags in the XLIFF file.

3 Related Work

Koehn and Senellart (2010) discuss the integration of SMT into translation memories. They also discuss XML markup, but not because they are interested in tag transfer from source to target. Instead, they use XML to mark untranslated parts of a source sentence, after most of it has been matched in the translation memory; the marked-up portions are passed to SMT.

Du et al. (2010) focus on the central topic of the current paper: how to handle markup in an SMT system whose output will be post-edited. They compare the performance of Moses (Koehn et al., 2007) using three different tag transfer

¹ The differences observed between our TMX and XLIFF files may be explained by the version of the software used to create them: SDL Trados 2007 for TMX files, SDL Trados Studio 2011 for XLIFF. However, our aim is to handle any TMX or XLIFF file submitted to our software.

methods on Symantec French and English data in TMX format: “Complete Tokenisation”, “Partial Tokenisation”, and “Markup Transformation”.

These methods lie along a spectrum between the one-stream and two-stream approaches. “Complete Tokenisation” is a pure one-stream method that treats tags as normal input, tokenizing them just like regular text. This leads to much longer sentences and diminishes the ability of the phrase table and the language model (LM) to learn useful patterns: the phrase table may have more noise in it because of bad word alignments, and the LM will be based on much sparser statistics. Worst of all, tags may be reordered during SMT, leading to syntactically incorrect TMX in the output, so a tidying post-processing stage is needed.

“Partial Tokenisation” has the same data flow as “Complete Tokenisation”, but handles tags specially. It deals with data sparsity by grouping frequent tag sequences into a single token, and also by grouping tags and tag sequences into categories, each assigned a single symbol. Thus, sentences are about 50% shorter than in “Complete Tokenisation”. This method can handle tags not seen in the training data.

“Markup Transformation” might be called a “1.7-stream” approach. As in the pure two-stream approach, markup is stripped from the input prior to SMT and then reinserted in the resulting translation. However, the decoder is not given complete freedom to reorder words: when markup is stripped from a region of input text, that region is marked as a Moses “zone”. During SMT, words may be reordered within the zone, and the zone can be moved around, but there may be no movement of words between the inside and the outside of the zone. Thus, no tricky cases can arise where word reordering causes words to cross tag region boundaries (see **Section 4.2** below); however, the translation quality may suffer.

The authors found that human evaluators do not have a strong preference for any one of these methods. Automatic metrics (NIST, BLEU, TER, and MTR), however, overwhelmingly prefer the two “Tokenisation” methods. There is some evidence that the “zone” restriction reduces the quality of the output.

Tezcan and Vandeghinste (2011) apply the one-stream approach to TMX data for English-to-French and English-to-Spanish data, also using Moses. They look at four methods. One of

these methods is a pure one-stream implementation roughly equivalent to “Complete Tokenisation” in Du et al. (2010). The other three methods can be seen as variants of “Partial Tokenisation” as described in Du et al. (2010), differing in the degree of generalization of tags and in the mechanisms for placing tags in the target text. The experiments in this paper seem to show a slight advantage for “Role-Based Markup Normalization”, in which tags are grouped into categories based on their roles, with each category given its own token. This improves the quality of the output compared to that from complete tokenization to an extent equivalent to the training data being increased by 50–100%.

Zhechev and van Genabith (2010) describe a system that matches chunks of the input sentence with chunks found in a translation memory where possible, then uses SMT decoding to fill in the unmatched pieces. Although tag transfer is not the main subject of the paper, the system described does transfer tags from the source to the target. As in the “Partial Tokenization” method above, tags are replaced by a simplified representation. Reading between the lines, one infers that the tag transfer approach here is probably “one-stream” with the simplified tags left in the training data, not stripped out prior to decoding.

The work closest in spirit to our own approach is described in Hudík and Ruopp (2011), as part of a description of the ongoing Moses for Localization (M4Loc) project undertaken by the open-source Moses community. The overall pipeline described by Hudík and Ruopp (2011) is identical to ours, but some important details differ. For instance, the authors use only phrase alignment information to transfer tags, but not word alignment information, inserting tags only at phrase boundaries in the target text, which can easily result in misplaced tags. In our system, we use the word alignment from the phrase table to correctly place tags within segments (in addition to phrase pair information). We handle both TMX and XLIFF formats, while they handle only XLIFF (a defensible decision, as XLIFF is replacing TMX). The software described in Hudík and Ruopp (2011) has been released (M4Loc, 2012), but has undergone some subsequent changes.

Finally, our work generalizes software mechanisms implemented by our colleague George Foster for transferring simple markup in Canadian parliamentary data, such as the political affili-

ations of people speaking in a debate, from one language to another (Foster et al., 2010).

4 Our Approach

4.1 Data Flow

We implemented two-stream tag transfer in our in-house system, a phrase-based SMT system resembling Moses (and which is licensed to translation agencies). As in Moses, the most frequent word alignment is stored with each phrase pair. **Figure 1** illustrates the data flow:

- **XML file:** This is the input file in XML format (either TMX or XLIFF).
- **Extract:** From XML files, extract the list of input sentences to translate, including their formatting tags. For XLIFF, tags are kept as is: `<g id="i"> words... </g>` for paired tags, `<x id="i"/>` for isolated tags. TMX tags are more complicated, so they are wrapped in `<open_wrap id="i">`, `<close_wrap id="i">` or `<tag_wrap>` tags that are designed to be parsed by simple regular expressions.
- **Tokenization:** We use our standard tokenizer, customized to be aware of XLIFF and wrapped TMX tags. We tokenize the actual text, not tags, but the tags inform some choices. E.g., the tokenizer normally splits tokens at tag boundaries, but if a token contains an open/close tag pair, we keep it together, so 31st, CO₂, etc., are kept as single tokens, despite the superscript and subscript tags they contain.
- **Q.tags.tok file:** This file contains tokenized text with the tags still embedded.
- **Strip tags:** By stripping out the tags, we get standard tokenized text of the kind our decoder normally works with.
- **Decoder pipeline:** This is our normal SMT pipeline: lowercase, apply rules (e.g., date and number parsing), decode, truecase.
- **P.tok file:** This is the decoder pipeline’s main output – tokenized truecased text.
- **P.trace file:** The decoder trace includes phrase segmentation and alignment from the decoding process, and word alignments associated with each phrase pair.

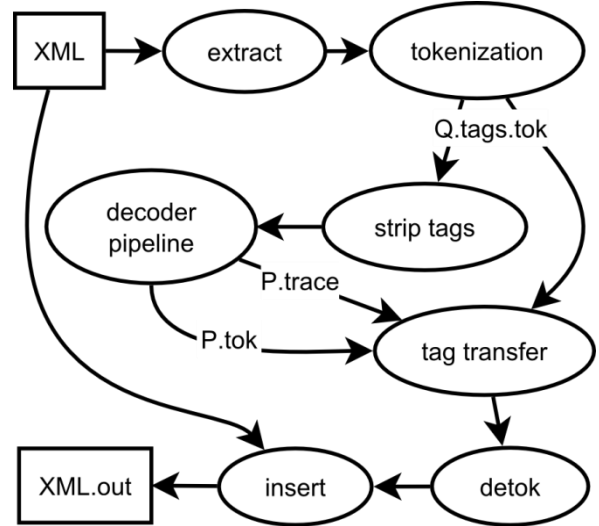


Figure 1. Two-Stream Data Flow.

- **Tag transfer:** This is the core module for this paper. It uses segmentation and word-alignment information from **P.trace** to insert the tags from **Q.tags.tok** correctly in **P.tok**.
- **Detok:** Detokenization is done using our standard detokenizer, again customized to be aware of tags.
- **Insert:** The translated detokenized sentences including their formatting tags are re-inserted into the original XML.
- **XML.out file:** This is the final TMX or XLIFF output file containing the source and translated text, including tags.

4.2 Tag transfer rules

The success of two-stream approaches depends on the **tag transfer** rules. These rules know about the phrase pair alignments used by the decoder to generate **P.tok** and the word alignments within each phrase pair.

Let “source tag region” (STR) refer to a contiguous sequence of source words that is enclosed by matched “open” and “close” tags, let “target covering phrases” (TCP) be the target-language phrases used to decode the source tag region, and let the “source phrase region” (SPR) be the words in the source covered by the source-language phrases corresponding to the TCP. The nature of decoding ensures that the SPR is a contiguous word sequence, but it may extend beyond STR, whereas TCP need not be contiguous.

The core transfer rules for paired tags are:

case 1: 1 STR, 1 phrase in SPR, STR=SPR, tgt phrases contiguous

Src = he returned home [**safe and sound**] . Phrase pairs: "safe and sound" || "sain et sauf"
 STR = SPR
 Tgt = il est retourné chez lui [**sain et sauf**] .

case 2: 1 STR, 2+ phrases in SPR, STR=SPR, tgt phrases contiguous

Src = he returned home [**safe and sound**] . Phrase pairs: "safe" || "sain", "and" || "et", "sound" || "sain"
 STR = SPR
 Tgt = il est retourné chez lui [**sain et sain**] .

case 3: 1 STR, STR≠SPR, tgt phrases contiguous

Src = he loves [**wine, women and song**] . Phrase pairs: "women and song" || "les femmes et les chansons", "loves wine," || "aime le vin,"
 STR
 SPR
 Tgt = il aime [**le vin, les femmes et les chansons**] .
 word alignments

Figure 2. Some Special Cases.

case 4: STR=SPR, tgt phrases not contiguous

Src = [**hang up**] the phone . Phrase pairs: "hang" || "lege", "up" || "auf"
 STR = SPR
 Tgt = lege das telefon auf .
 ? [**lege** das telefon **auf**] or [**lege**] das telefon [**auf**] or lege das telefon [**auf**] ?

OUR RULE PICKS THIS

case 5: disjoint STRs yield overlap in target

Src = talk [**at noon**] **tuesday** [**with bob**] **at my place** . Phrase pairs: "at noon" || "à midi", "tuesday" || "ce mardi", "with bob" || "avec bob", "at my place" || "chez moi"
 STR₁ STR₂
 Tgt = parler [**à midi**] **avec bob** **ce mardi** [**chez moi**] .
 TCP₁ TCP₂

IN TMX FILES WE CAN KEEP THE CROSS-OVER:

Tgt = parler [**à midi**] **avec bob** **ce mardi** [**chez moi**] .

OUR XLIFF RULE MOVES THE LEFT "CLOSE" TAG RIGHTWARD:

Tgt = parler [**à midi**] **avec bob** **ce mardi** , **chez moi** [] .

Figure 3. More Complex Cases.

1. If the boundaries of the STR and the SPR coincide exactly, and the TCP are contiguous in the output target-language sentence, then the tag pair surrounding the STR is copied into the output so it surrounds the TCP. This rule is shown in **Figure 2**, **cases 1** and **2**. In **case 2**, note that phrase reordering during translation does not affect application of the rule.
2. If the boundaries of the SPR extend beyond the STR, even if the TCP are contiguous in the output target-language sentence, there may be target words from the TCP that are word-aligned with source words outside the STR. This is shown in **Figure 2**, **case 3**: "aime" comes from the phrase pair ("loves wine ,", "aime le vin ,") whose source side is partly inside, partly outside the STR; "aime" is a word aligned with a word inside the SPR ("loves") but outside the STR. This is where word-alignment information helps us. Here, we copy the tag pair surrounding the STR into the output so that it surrounds every target word that is word-aligned with a word in the STR. So "aime" is not **bold** even though it is in a phrase that is partially bold.
3. If the TCP are not contiguous in the output, we copy the tag pair from the STR into the output in such a way that all target phrases in the TCP are surrounded by the tag pair, thus extending the tag to apply to intervening phrases too. (We may also apply rule 2 if the SPR extends beyond the STR, to decide which target words to include in the phrases at the boundaries of the TCP). **Figure 3**, **case 4** shows this along with two alternative rules that could have been applied.

4. Sometimes, two or more disjoint STRs may generate overlapping TCPs, as in **Figure 3**, **case 5**. Here, STR₁ applies Lucida Handwriting typeface to words in its scope, STR₂ applies bold Times New Roman typeface, and the TCPs overlap. The above rules will yield tags that cross over in the output. This is seldom desirable, but there is no good rule to fix it automatically. TMX format allows its <bpt> and <ept> tag pairs to cross over, so we leave this situation as is in TMX files, letting the translator fix it by hand. However, XLIFF does not permit cross-over of its <g>/</g> tag pairs, so we must remove the cross-overs from the output. We arbitrarily chose to move the first close tag past the second close tag, creating a nested but legal structure for post-edition.

The rules for transferring isolated tags are:

1. If an isolated tag is a point tag in the document, it is considered to be attached to the next source word, and is placed before the target word aligned to that source word.
2. If an isolated tag is the close or open tag of a tag pair that starts or ends outside the current segment in the underlying document, then it is treated using rules 1 to 3 above, as if it was paired with an open tag at the beginning of the segment, or a close tag at the end of the segment, respectively.

In addition to these core rules, some care is taken to preserve the source order when multiple tags end up between the same two target-language words. Note that both cases in **Figure 3** might have been handled better by a one-stream system, which could have kept target words orig-

inating from the same STR together. However, this might have made the translation itself worse.

5 Mini-Evaluation

We manually evaluated the accuracy of tag placement in the output of our system on one document. This is a “mini-evaluation” because it involved only one document, and it did not look at post-editing effort, just tag placement. The goal was to understand what’s going on, rather than to draw statistically defensible conclusions.

We trained our in-house phrase-based SMT system on 384K English-French sentence pairs collected during Nov. 2007 – Feb. 2008 from the Environment Canada (EC) web site of Environment Canada (www.ec.gc.ca), with 5.1M English and 6.1M French words in total. The EC website constitutes a nice parallel corpus because it has varied types of formatted content published in French and English simultaneously. Our system obtains BLEU scores in the low fifties on randomly sampled dev and test sets of 1500 sentences (trained/tested on text stripped of tags; one reference for dev and test).

We tested tag transfer on a more recent EC document, cutting/pasting the HTML text into MS Word. The TMX test file was created by extracting segments from the document with SDL Trados 2007. It had 367 segments, most with multiple sentences. There were 298 open/close tag pairs and 161 isolated tags for a total of 757 tags, with on average 2.1 tags per segment (tag pairs count as two tags). 56% of segments had no tags, 22% had one tag, while the remaining 22% had two or more; 6% had ten or more. The segment with the most tags had 24.

The XLIFF test file was created by extracting the segments from the document with SDL Trados Studio 2011. It had 1026 segments, most containing a single sentence. There were 194 open/close tag pairs and only 2 isolated tags for a total of 390 tags, with an average of 0.38 tag per segment. 91% of the segments had no tags, 3% had two tags (one tag pair), 6% had more; only 0.6% of segments had more than ten. The highest tag count for a segment was 14.

The TMX and XLIFF statistics are not strictly comparable because the XLIFF data are segmented into sentences, unlike the TMX data. However, there are fewer tags in the XLIFF, which makes tag transfer easier. In particular, the near absence of isolated tags in XLIFF shows

that it abstracts more formatting information out of the segments.

We gave each tag in each segment produced by our system one of the following annotations (we evaluated only tag placement, not translation quality): **Good** (tag is well placed); **Reasonable** (tag needs to be moved, but is next to text that also needs to be moved, so these can be moved together); **Wrong** (tag is wrongly placed); **No chance** (the decoder’s output is so bad there is no correct place for the tag). We added **Spacing error** if space before or after the tag should be deleted or inserted.

	TMX	XLIFF
Total no. of tags	757	390
Good	704 (93%)	351 (90%)
Reasonable	23 (3%)	8 (2%)
Wrong	27 (4%)	30 (8%)
“No chance”	3 (.4%)	1 (.3%)
Spacing error	165 (22%)	59 (15%)

Table 1. Mini-evaluation results by tags.

Table 1 shows tag-by-tag results. Most tags are placed correctly, reducing post-editing effort compared to a system with no tag transfer. A number of reasonable or wrong tags still need to be moved, though most of the remaining post-editing will fix spacing before or after tags. Scoring of spacing errors was strict: e.g., a space moved from before a tag to after it was counted as two errors (1 deletion + 1 insertion). Some of these spacing issues might not matter in practice.

	TMX	XLIFF
Segments with at least one tag	161	88
All tags good, no spacing errors	88 (55%)	33 (38%)
All tags good	143 (89%)	66 (75%)
All tags good or reasonable	144 (89%)	67 (76%)
At least one tag wrong or no chance	17 (11%)	21 (24%)
At least one spacing error	70 (43%)	41 (47%)

Table 2. Evaluation results by segments.

Table 2 shows our results by whole segments. Many segments have all tags placed correctly (89% / 75%) and a lower but still large proportion have completely correct spacing around tags (55% / 38%), and thus require no post-editing.

We also tried two methods for obtaining the word alignment within phrase pairs. The first,

“heuristic word alignment”, uses heuristics developed by Foster et al. (2010) to recover a possible word alignment for a translation, given only its phrase pairs; the heuristics employ cognate information combined with a bias towards linear alignment. The second, “original word alignment” applies the original word alignment used to create the phrase table and stored in it. The two methods produce almost the same output; the number of differences is too small to draw any firm conclusions. Although a software bug affected some of our results, it is clear that the two methods perform at about the same level. This may be because the word alignment is not often required to place tags correctly, or perhaps because the heuristics of Foster et al. (2010) are well suited to the task at hand.

6 Discussion

We thought of, but did not try, some intriguing alternative rules for tricky tag transfer situations. In future, we’d like to try:

- Biasing the system to favour the original tag order whenever possible. Our current rules only require an “open” tag to precede the matching “close” tag, and require the XML in the target to be valid.
- An extreme version of this: a hard rule that considers contiguous sequences of tags as if they were just one tag. The translators/post-editors we observed working in SDL Trados TagEditor seemed to do this, and shortcuts in the interface support it.
- Better rules for handling spacing. During the mini-evaluation, we noticed that the detokenizer often left unwanted spaces around tag sequences. This may be due to a design error: the tag transfer module makes spacing decisions based on tokenized input. Then the detokenizer removes unnecessary spaces, but without knowledge of the source text. The system could use raw input, before tokenization, to decide what spaces need to be kept.

The major limitation of our paper is the lack of an experimental comparison between our method and alternative ones on a real post-editing task. A thorough study would compare a variety of one-stream and two-stream methods from the literature. It could employ human evaluation as in Du et al. (2010), or follow the more extensive proto-

col of Green et al. (2013), using eye, cursor and mouse tracking to measure the impact of different methods on post-editing. With respect to automatic metrics, evaluation of one-stream methods should involve at least two scores, one comparing tag-free output with tag-free references and one comparing tagged output with tagged references, to distinguish the impact of the methods on translation quality and their impact on formatting. Pure two-stream methods only affect formatting, so only the second score is necessary.

Another limitation is that we studied a language pair with similar word order (English-French). Language pairs with more reordering (e.g., German-English) make it harder for the pure two-stream approach to make sensible tag transfer decisions. Language pairs with different writing systems (e.g., Arabic-English or Chinese-English) pose deeper questions: e.g., what are the semantics of **bold** or *italic* in languages not written in the Roman alphabet? One might even wish to explore whether capitalization has semantic equivalents in other writing systems.

This paper describes implementation of a pure two-stream approach for transferring tags from source to target text in an SMT system. “Two-stream” means that tags are stripped from source text before the decoder sees them, and transferred along a separate route, as opposed to the “one-stream” approach where the decoder handles the tags. Our approach gives the decoder more freedom, but requires complex tag reinsertion rules. Experimental studies of both approaches in a realistic post-editing situation will be required to decide which approach is better.

Acknowledgement

We would like to thank CLS Lexi-tech Ltd. for creating our test TMX and XLIFF files. We would also like to thank the reviewers for their constructive suggestions for improving this paper.

References

- Du, Jinhua, Johann Roturier and Andy Way. 2010. TMX Markup: A Challenge When Adapting SMT to the Localisation Environment. *EAMT 2010*, Saint-Raphaël, France, 253-260.
- Foster, George, Pierre Isabelle and Roland Kuhn. 2010. Translating Structured Documents. *Association for Machine Translation in the Americas (AMTA 2010)*, Denver, Colorado, USA.
- Green, Spence, Jeffrey Heer, and Christopher Manning. 2013. The Efficacy of Human Post-Editing

- for Language Translation. *CHI '13, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
- Hudík, Tomáš and Achim Ruopp. 2011. The Integration of Moses into Localization Industry. *Proceedings of EAMT 2011*, Leuven, Belgium.
- Koehn, Philipp and Jean Senellart. 2010. Convergence of Translation Memory and Statistical Machine Translation. *AMTA Workshop on MT Research and the Translation Industry*, Denver, Colorado, USA.
- Koehn, Philipp, Hieu Hoang, *et al.* 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, Czech Republic.
- Localization Industry Standards Association (LISA). 2005. TMX 1.4b Specification. Accessed May 17, 2013. <http://www.gala-global.org/oscarStandards/tmx/tmx14b.html>
- M4Loc. 2012. m4loc – Moses for Localization. Accessed May 14, 2013. <https://code.google.com/p/m4loc/>
- OASIS. 2008. XLIFF Version 1.2 Specification. Accessed May 17, 2013. <http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html>
- SDL International. 2011. SDL Trados Studio 2011 Languages and Filters (guide). <http://www.translationzone.com/en/landing/premium-downloads/sdl-trados-studio-2011-languages-and-filters.asp>
- Tezcan, Arda and Vincent Vandeghinste. 2011. SMT-CAT integration in a Technical Domain: Handling XML Markup Using Pre & Post-processing Methods. *Proceedings of EAMT 2011*, Leuven, Belgium.
- Zhechev, Ventsislav and Josef van Genabith. 2010. Seeding Statistical Machine Translation with Translation Memory Output through Tree-Based Structural Alignment. *4th Workshop on Syntax and Structure in Statistical Translation (SSST-4)*, Beijing, China.

Assessing Post-Editing Efficiency in a Realistic Translation Environment

Samuel Lübli¹ Mark Fishel¹ Gary Massey² Maureen Ehrensberger-Dow² Martin Volk¹

¹ Institute of Computational Linguistics ² Institute of Translation and Interpreting
University of Zurich Zurich University of Applied Sciences
Binzmühlestrasse 14 Theaterstrasse 15c
CH-8050 Zürich CH-8401 Winterthur

{laeubli, fishel, volk}@cl.uzh.ch {mssy, ehre}@zhaw.ch

Abstract

In many experimental studies on assessing post-editing efficiency, idiosyncratic user interfaces isolate translators from translation aids that are available to them in their daily work. In contrast, our experimental design allows translators to use a well-known translator workbench for both conventional translation and post-editing. We find that post-editing reduces translation time significantly, although considerably less than reported in isolated experiments, and argue that overall assessments of post-editing efficiency should be based on a realistic translation environment.

1 Introduction

Machine translation has had a considerable impact on the translation industry in recent years, and there are a number of studies that systematically test the efficiency of replacing manual translation “from scratch” with post-editing—the process of manually adapting machine translations. Most of such studies isolate participants from access to additional translation tools since they would affect the timing and other response variables. To ensure precise measurements, translators usually operate via a simple user interface, specially tailored for such studies.

This means, however, that the conditions under which some studies are conducted differ from the final working conditions in which post-editing will be used. The interfaces used are not directly comparable to the translation aids and workbenches currently available to translators in their daily work, which can result in overestimations of

time gains through post-editing. We believe that assessments of post-editing efficiency should instead be based on a more *realistic environment* for participating translators.

In this paper, we employ a customary translation workbench to evaluate the effectiveness of post-editing translations of marketing texts from the automobile industry. We hypothesize that providing translators with a domain-specific translation system will increase their productivity even when added on top of other well-known translation aids, such as translation memories and bilingual terminology databases. In line with recent work on inferential statistics in post-editing research by Green et al. (2013), we test whether this productivity increase is statistically significant.

In the following section, we outline our use case and briefly review a number of post-editing studies that have been conducted so far. In Section 3, we detail our experimental design, outlining how we measure translation throughput with and without post-editing (Section 4), as well as the quality of all translations produced in the study (Section 5). In Section 6, we compare the respective results to related studies and contrast them with user perceptions, and then we draw conclusions and outline future work in Section 7.

2 Background

The research reported here is part of a joint project between the University of Zurich and a language service provider (LSP) with a primary focus on translating material from the automobile industry, such as brochures and other marketing texts—a specific domain with its own terminology and typical translations (see Table 1). The aim of the

Source (DE)	Reference (FR)	TM-Only (FR, P4)	Post-Edit (FR, P6)	English gloss
Streifenbeklebung auf Frontklappe und über den Seitenschwellern	Bande adhésive sur le capot et sur les seuils latéraux	Bandes autocollantes sur le capot et sur les jupes latérales	Bandes adhésives sur le capot et au niveau des ailes	<i>Adhesive stripes on the bonnet and above the side skirts</i>

Table 1: A German source segment in translation task D (product features) and its translations into French: A reference translation produced by a specialized translator prior to our experiment and translations by participants in the TM-Only and Post-Edit conditions as well as the English gloss.

project is to build a domain-specific machine translation system for the LSP to use in a post-editing scenario.

There have been a number of studies that assess the efficiency of post-editing (e.g., Guerberof, 2009; Sousa et al., 2011; Green et al., 2013). Most of these set up controlled environments for their experiments and develop specially tailored user interfaces for post-editing tasks (e.g., Aziz et al., 2012). The main reason for this is the priority placed on precise measurements of translation time, pause durations and input device activities.

Some evaluations of post-editing in an industrial context have also been reported. For example, Plitt and Masselot (2010) replace manual translation with post-editing for software localization. Other industry-oriented studies (Volk et al., 2010; Flournoy, 2011) focus more on the challenges of deploying machine translation in the respective sector or company.

Our work strives to combine the key elements of these two approaches: We ensure precise time and activity measurements while preserving a realistic translation environment.

3 Experimental Design

The main idea behind our experiments is to assess the efficiency of post-editing in a realistic translation environment. The participating translators were asked to translate a number of texts in two conditions using Across Personal Edition¹. The setup for the TM-Only condition included access to a translation memory (TM) with 176,957 domain-specific entries. Exact matches of the TM were automatically inserted into the otherwise empty translation template, and fuzzy matches were displayed in a dedicated section of the workbench. In addition, the participants were able to access a small domain-specific terminology database

(704 entries) as well as any additional translation aids of their choice, such as printed or online dictionaries.

In the Post-Edit condition, machine-translated output was included in addition to the previously described setup, while access to the same translation aids was allowed. However, whereas in the TM-Only condition text fields with no exact match in the TM were left empty, they were filled with machine translations (MT) in the Post-Edit condition. Machine-translated segments were marked as such, so that the origin of the translation was transparent to the translators. The participants were asked to translate the German source text (TM-Only) or revise the French MT output (Post-Edit) as needed to produce high-quality French target texts. They were encouraged to work however they wanted and had access to the fuzzy TM matches, the terminology database, and online resources.

Pre-edit translation drafts were produced by a domain-specific statistical machine translation system. It was built using the same translation memory data that was used for the present study, as well as out-of-domain parallel corpora; a more detailed description can be found in Läubli et al. (2013a,b). We implemented a simple RPC-based software link to enable seamless integration of the translation system into the translation workbench.

The German source texts for the translation tasks were provided by the LSP. We selected four typical texts (A–D) that cover specific aspects of our domain in scope (see Table 2). Since all of the texts had been translated by professional translators at the LSP in their normal workflow, we also had access to the corresponding reference translations into French. This allowed us to compare the output of translators experienced in the automobile industry text domain (the LSP staff) with that of those new to it (participants of this study), as well as to assess the effect of post-editing on this comparison.

¹<http://www.my-across.net/en/translation-workbench.aspx>

Text	A	B	C	D
Type	company portrait	letter	presentation slide	product features
Language	full sentences	full sentences	bullet points	bullet points
Segments	7	18	12	13
Words	107	103	50	64
Characters	890	742	489	557
Coverage	poor	good	poor	good
100%	-	3	1	3
80–99%	-	6	-	5
50–79%	1	3	1	2
No match	6	6	10	3

Table 2: Text materials. We used four typical marketing texts from the automobile industry.

The six participants of the present study (P1–P6) were native speakers of the target language (French) and highly competent in the source language (German), between 20 and 40 years of age (mean: 25.5, median: 22.5). All of them were familiar with translation workbench technology and were majoring in German-French translation in a BA program in general translation or in an MA program in specialized translation. Four participants reported regularly translating texts for payment, and two of them had already been employed as professional translators. The participants were compensated for their involvement in the study according to customary rates at their home institute.

Each of the six participants (P1–P6) translated all four texts (A–D) after a familiarization session with the Across translation workbench system. Participant-document assignment was done randomly with three constraints, to guarantee that:

- (i) no participant was presented with the same document twice, in any setup;
- (ii) each document was translated three times in TM-Only and three times in Post-Edit;
- (iii) each participant translated two documents in each condition.

The time needed for completing the translation tasks was measured by means of screen recordings. This technique, commonly used in transla-

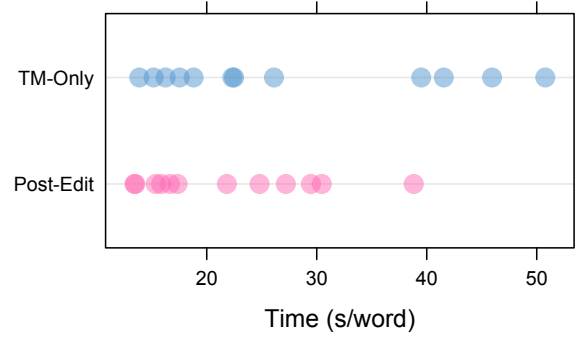


Figure 1: Translation time (seconds per word) by condition.

tion process research (e.g., Ehrensberger-Dow and Massey, 2013), clearly distinguishes our experimental design from previous studies as the use of screen recordings allows precise time measurements to be made without the need for idiosyncratic user interfaces (such as Aziz et al., 2012) or relying solely on what participants themselves track and report (as, for example, in Plitt and Mas-selot, 2010).

In the following, we describe the results of efficiency estimation experiments (Section 4), before assessing the quality of the translations produced in the two conditions (Section 5).

4 Translation Throughput

As mentioned above, no document was translated more than once by the same participant. Although this guarantees independent time measurements (translating a document in either condition would have inevitably affected translation timing in the other condition), this also means that we cannot compare those time measurements directly due to several variables such as average translation speed per participant, document length and complexity, etc. Instead, we normalized the measurements.

The standard approach is to normalize the time by the length of the document: i.e., to divide the time per translation task by the number of sentences or words in its source text. However, this only accounts for the length of the document and not for other text characteristics that are more difficult to control in an experimental setting (cf. Table 2). In our data, length-normalized translation times vary considerably by document (see Ta-

Text	Time (s/word)		Change
	TM-Only	Post-Edit	
A	16.0 ± 2.6	16.2 ± 1.0	1.5%
B	18.8 ± 3.1	14.5 ± 1.8	-22.6%
C	45.4 ± 5.7	31.4 ± 7.1	-30.9%
D	30.0 ± 10.2	26.2 ± 3.9	-12.8%

Table 3: Average translation time and standard deviation (seconds per word) by document. Each document was translated by three randomly assigned participants per condition.

ble 3), indicating that other factors² influence the average time needed for translating a word.

Overall, translating a word took participants 27.5 seconds in TM-Only and 22.1 seconds in Post-Edit on average (-19.9%). Standard deviations are high in both conditions (see Figure 1), but clearly higher for TM-Only (± 13.2 seconds) than for Post-Edit (± 8.1 seconds). Translation speed differs greatly between both participants and documents (see Figure 2). The more prose-like texts consisting primarily of full sentences (A, B) were translated much faster than the information-denser texts consisting primarily of bullet points (C, D), regardless of whether the TM coverage was good or poor.

According to the length-normalized measurements, post-editing helped four out of six participants translate faster. However, performing a *per subject* analysis is not appropriate in our setting as each participant translated two texts per condition and individual averages per participant and condition are based on two time measurements only. Results of a *per item* analysis (Table 3) show that three out of four texts were translated faster in Post-Edit; however, the same criticism applies to these results, since time measurements are averaged over only three participants per condition and document.

Looking at our data by participants *or* items is not satisfactory. As we seek to generalize from samples to populations—ideally, all possible translators rather than P1–P6 and all automobile marketing texts rather than texts A–D—we are inter-

²For example, Green et al. (2013) report a significant correlation between translation time and percentage of nouns in the source text. In our experiment, participants also required more time to translate texts written in nominal style, i.e., bullet points rather than full sentences.

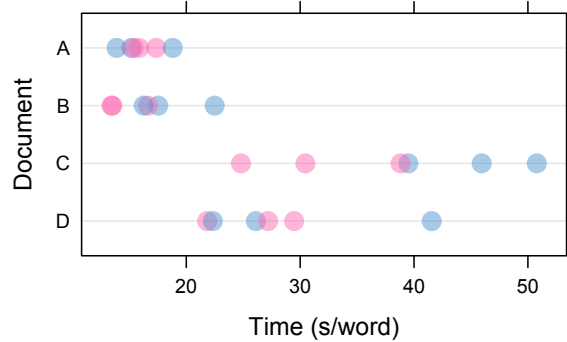


Figure 2: Translation time (seconds per word) by document. Blue and pink data points represent TM-Only and Post-Edit, respectively.

ested in assessing whether Post-Edit is faster than TM-Only given random variation from both participants and documents. We thus want to test for a genuine difference between our conditions despite extraneous or potentially confounding variables that we cannot fully control in our experiment, such as different translation speed between participants or the frequency of certain word classes in texts.

We thus employed linear mixed effects (LME) models (Baayen et al., 2008) to analyze our data.³ Green et al. (2013) showed that LME models are preferable to, e.g., analysis of variance (ANOVA) in post-editing experiments because language can be treated as a random effect, thus avoiding the “language-as-fixed-effect fallacy” (Clark, 1973). Accordingly, we only use the translation condition as a fixed effect and both participant and text as random effects. We did not apply a prior normalization of times by text length since length is an implicit feature of the respective random effect. We checked for homogeneity and normality in our data and tested the validity of the mixed effects analyses by comparing the models with fixed effects to null models (comprising only the random effects) through likelihood ratio tests.

The resulting LME model shows a significant main effect for translation condition (MCMC-estimated p -value = 0.0192). The estimated average translation times are 1,957.4 seconds per text for TM-Only and 1,617.7 seconds per text for Post-Edit; i.e., post-editing reduces time by 17.4%.

³Using the `lme4` (Bates et al., 2013) and `languageR` (Baayen, 2011) packages in R (R Core Team, 2013).

5 Translation Quality

Time savings through post-editing are only relevant if the quality of the produced translations remains consistent. We consider this criterion to be met if the target audience cannot distinguish post-edited from conventionally translated texts, which we tested with translation experts (Section 5.1) and with the participants of our study (Section 5.2).

5.1 Expert Rating

To assess overall translation quality, we asked two independent experts to evaluate all translations in detail. We included the reference translation provided by the LSP, resulting in seven translations per source text—six produced by P1–P6 (three in TM-Only, three in Post-Edit) and one produced by a professional translator. The experts were not informed about the origins of the translations or the translation conditions. They were provided with the four German source texts (A–D) and the seven French translations of each in unified formatting and random order. Both experts returned the assessments of the 28 translations within a week of receiving them and then had no further involvement in the study. They were compensated according to regulations of the Zurich University of Applied Sciences (ZHAW).

The experts, lecturers in German-French Translation, used the ZHAW’s internal evaluation scheme, consisting of five ordinal scales for (i) target language expression, (ii) target language grammar, (iii) target language syntax, (iv) semantic accuracy, and (v) translation strategy. For every translation that the experts rated, they were asked to score each main category on a scale of 1 to 4 points each; the higher the score, the better the performance in that category was considered to be. The five separate scores were aggregated to obtain a value out of 20 for each translation.

The averaged expert ratings reveal clear differences between the TM-Only and Post-Edit conditions for the translations of the texts containing fully formed sentences (A, B; see Table 4). With an average total score of 15.7, the post-edited translations of text A were rated 16% higher than the TM-Only target texts, which scored only 13.5 on average. The post-edited translations of text B scored 15.0, 7% higher than the average total (14.0) for the TM-Only condition. The distinctions between the conditions for texts C

Text	Expert Rating			Diff. ^a
	TM-Only	Reference	Post-Edit	
A	13.5 ± 1.4	14.0	15.7 ± 1.9	16.3%
B	14.0 ± 1.9	13.5	15.0 ± 1.0	7.1%
C	13.8 ± 1.3	15.5	14.2 ± 1.9	2.4%
D	16.0 ± 1.7	15.5	15.7 ± 1.1	-2.1%

^a Difference between TM-Only and Post-Edit averages.

Table 4: Expert ratings (points on an ordinal scale; max=20 points). Scores for TM-Only and Post-Edit are averages over two independent ratings each for three translations per condition and document. Reference scores are averages of two ratings for one professionally produced translation per document.

(TM-Only: 13.8, Post-Edit: 14.2) and D (TM-Only: 16.0, Post-Edit: 15.7) are less (2.4% and -2.1% respectively), conceivably because coherence is less of an issue in texts almost exclusively made up of bullet points. It would seem that the Post-Edit condition produced full-sentence texts of higher quality.

This appears to be confirmed by the expert scores for the reference translations. Compared with the translations produced in Post-Edit, the reference translations received lower average ratings for the full-sentence texts A (company portrait) and B (letter). In two cases, namely texts B and D, the reference translations also scored worse than some of the target texts written in the TM-Only condition.

5.2 Pairwise Ranking

In addition, we tested whether the participants (P1–P6) prefer professional translations produced by the LSP staff over those produced in the study. We applied a pairwise ranking procedure⁴ in which evaluators compare two translations $\langle t_1, t_2 \rangle$ of a given segment in the source language and choose the better fit, with ties allowed. We used six random German segments from each text (A–D) and had all participants compare the corresponding professional translation to those produced by all

⁴A similar procedure was used at the 2012 Workshop for Machine Translation (Callison-Burch et al., 2012). In contrast to other human evaluation metrics such as fluency and adequacy judgments on ordinal scales, pairwise rankings are usually more comprehensible and better reproducible for non-expert evaluators.

Condition	Wins	Ties	Losses	<i>p</i> -value
TM-Only	112	94	154	0.012
Post-Edit	128	96	136	0.667

Table 5: Pairwise ranking of translations produced in the study against professional reference translations. *p*-values indicate genuine differences between the number of wins and losses (Sign Test).

other participants, such that each participant evaluated 120 $\langle t_{\text{professional}}, t_{\text{participant}_i} \rangle$ tuples in total. Participants were not told about the origin of the translations to be compared, i.e., they did not know that they were comparing professional translations to those produced in the study. We presented the translation alternatives in random order and inserted 10 “spam” items per participant—tuples where one translation did not match the original segment—to control for deliberate choices.

Results of the ranking are presented in Table 5. When comparing the LSP’s translations to those that have been produced in the TM-Only condition, participants preferred the former: The reference translations were preferred in 154 out of 266 non-tie cases (57.9%); the translations by P1–P6 in 112 cases (42.1%). In contrast, reference and participants’ translations were rated comparably in the Post-Edit condition: The former were preferred in 51.5% of the cases, the latter in 48.5%.

We applied the Sign Test to determine whether the win:loss ratio between TM-Only and the professionally produced reference translations (hereafter “Reference”) as well as that between Post-Edit and Reference is genuine. As presented in Table 5, TM-Only is ranked significantly lower than Reference, while the difference between Post-Edit and Reference is attributable to chance. In other words, participants could not distinguish their post-edited translations from the professionally produced translations, while they considered the professional translations better than those produced in the TM-Only condition.

6 Discussion

Post-editing reduces time significantly even when a fully-featured translation workbench is available. Our results suggest that actual time savings lie within a range of 15–20%, which is, however, con-

siderably lower than numbers reported in other studies. For example, Sousa et al. (2011) found a “speeding-up [of] the translation process by 40%” for film subtitles from English to Portuguese. Plitt and Masselot (2010) report average time savings of 43% in software localisation from English to French, Italian, German, and Spanish.

One reason for this considerable difference may be that we did not recruit professional translators for our study. When compared to the results of Plitt and Masselot, who employed specialist translators with considerable experience in software localisation, the student participants P1–P6 needed a relatively long time to translate in both conditions (see Section 4 and Table 3). However, Sousa et al. obtained time savings of 40% even with volunteers that only “have some experience with translation tasks”. In contrast, our participants are pursuing and/or have completed university degrees in professional translation, and most of them regularly translate texts for payment (see Section 3).

In addition to other factors such as text genre and language pairs, the realistic translation environment might explain our high per-word translation time as well as the lower productivity gains. In contrast to many other studies (see Section 2), our translators were not forced to translate texts strictly segment by segment, since they were presented with a complete source text for each task rather than isolated sentences, and translated documents could be revised as a whole before submission. Inspections of screen recordings reveal that participants made extensive use of this possibility, which is common practice among professional translators (Guerberof, 2013). Most importantly, the availability of a domain-specific translation memory and a bilingual terminology database reduced the difference between TM-Only and Post-Edit, i.e., it increased translation throughput, especially in the former condition, where no machine translations were available.

Finding reasons for *why* translators still work significantly faster in the Post-Edit condition was not the focus of our study. However, a preliminary analysis of screen recordings supports Green et al.’s (2013) finding that translators draft less when post-editing. We also noticed that participants often neglected suitable fuzzy translation memory matches in the TM-Only condition. When the same or very similar translations

were automatically inserted into the target language template in *Post-Edit*, participants often accepted them with no or only minor changes. It seems that machine translations help translators by providing a clear “starting point”, thus eliminating the need for browsing through all available resources (translation memories, websites, or dictionaries) before actually starting to draft.

The evaluation of translation quality (see Section 5) confirms that post-edited translations are at least equivalent to conventionally produced translations. It also highlights the importance of considering the genre, information density, and linguistic structures of source texts when comparing the efficiency of various translation aids. Prose-like texts, such as company profiles and letters, may not be translated faster with MT input, but the final result may be of better quality overall because the translator can focus on editing the text to suit its purpose rather than focusing on translating words and structures. On the other hand, it takes much longer to translate information-dense texts (such as those consisting primarily of bullet points) from scratch, which is why good-quality MT input can help so much (up to 30.9%; see Table 3). With these types of informative texts, editing for cohesion and linguistic style is much less important.

The status of reference translations has been called into question by the findings reported in Section 5.1: Expert markers unaware of the translators’ background or training consistently categorized the reference translation as average quality compared with students’ translations. However, the comparison is somewhat unfair, since the conditions for the translations were not the same. Professional translators are subject to many constraints, such as time pressure and adherence to clients’ style guides, which were not imposed in the present study. From this point of view, a better design for the pairwise ranking might have been to compare the *TM-Only* and *Post-Edit* segments not only to the reference translation but also to each other.

A survey that followed the translation tasks revealed that our participants were considerably reserved towards machine translation and post-editing. Five out of six participants considered pre-translations in the *Post-Edit* condition to be “not useful” (4) or “not at all useful” (1); only one participant found them “sometimes use-

ful”. Overall, five participants preferred to work in the *TM-Only* condition, while one (P5) preferred *Post-Edit*, stating

For the very technical parts of the catalogue [to be translated] I would probably prefer the mode with pre-translations.

P3 indicated that the machine translations were helpful for translating difficult texts in terms of vocabulary,

[...] but I think [for translating] the two easiest texts [...], the pre-translations would have only confused me.

This stands in sharp contrast to the fact that post-editing resulted in significantly faster and even slightly better translations in our study. However, the discrepancy between translators’ perceptions and post-editing performance is a well-known phenomenon in the field (see, e.g., Koponen, 2012). On the other hand, our participants were by no means technology-averse in general: All of them used various computer-aided translation tools in the tasks and deemed both the domain-specific translation memory and the bilingual terminology database as either “sometimes useful” (3/3), “useful” (1/2), or “very useful” (2/1).

7 Conclusion

We have proposed a design for translation efficiency experiments that compares post-editing to computer-aided translation using a fully-featured translation workbench. In contrast to the simplified user interfaces deployed in other studies (e.g., Sousa et al., 2011; Green et al., 2013), this allows participants to use translation memories and terminology databases, long indispensable tools for professional translators. Precise time measurements were obtained by means of screen recordings, which is unobtrusive to participants and eliminates the need for them to track and report times themselves (as in Plitt and Masselot, 2010).

Applying the proposed methodology in a controlled experiment, we have shown that post-editing results in significantly faster translation with consistent quality even when compared to computer-aided translation (as opposed to completely unaided translation). While time savings are most noticeable in dense documents consisting of bullet points, post-editing also facilitated the

translation of prose-like texts that require editing for cohesion and linguistic style.

Specialist ratings as well as a pairwise ranking procedure confirm that the quality of post-edited texts is consistent with or, in some cases, even better than conventionally produced translations. Quality improvements through post-editing were mostly found in coherent full-sentence texts.

Overall, our results indicate that gains in translation throughput are around 15–20%, which is considerably lower than numbers reported in studies that isolate participants from commonly used translation tools such as translation memories and bilingual terminology databases. While such isolated studies are clearly important for examining specific aspects of post-editing, our findings strongly suggest that its overall efficiency should be assessed in a realistic environment that takes account of the various aids available to translators in their daily work.

Acknowledgements

We would like to thank Annina Meyer for her substantial help in organizing the quantitative experiments described in this paper. Our work was partly supported by Grant No. 11926.2 PFES-ES from the Swiss Federal Commission for Technology and Innovation.

References

- Wilker Aziz, Sheila Castilho, and Lucia Specia. Pet: A tool for post-editing and assessing machine translation. In *Proceedings of LREC*, Istanbul, Turkey, 2012.
- R. Harald Baayen. *languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics"*, 2011. URL <http://CRAN.R-project.org/package=languageR>. R package version 1.4.
- R. Harald Baayen, Douglas J. Davidson, and Douglas M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412, 2008.
- Douglas Bates, Martin Maechler, and Ben Bolker. *lme4: Linear mixed-effects models using Eigen and syntax*, 2013. URL <http://CRAN.R-project.org/package=lme4>. R package version 0.999999-2.
- Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of WMT*, pages 10–51, Montréal, Canada, June 2012.
- Herbert H. Clark. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4):335 – 359, 1973.
- Maureen Ehrensberger-Dow and Gary Massey. Indicators of translation competence: Translators' self-concepts and the translation of titles. *Journal of Writing Research*, 5:103–131, 2013.
- Raymond Flournoy. MT use within the enterprise: Encouraging adoption via a unified MT API. In *Proceedings of MT Summit XIII*, pages 234–238, Xiamen, China, 2011.
- Spence Green, Jeffrey Heer, and Christopher D. Manning. The efficacy of human post-editing for language translation. In *Proceedings of CHI*, Paris, France, 2013.
- Ana Guerberof. Productivity and quality in the post-editing of outputs from translation memories and machine translation. *International Journal of Localisation*, 7(1):11–21, 2009.
- Ana Guerberof. What do professional translators think about post-editing? *Journal of Specialised Translation*, 19:75–95, 2013.
- Maarit Koponen. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of WMT*, pages 181–190, Montréal, Canada, 2012.
- Samuel Läubli, Mark Fishel, Martin Volk, and Manuela Weibel. Combining domain-specific translation memories with general-domain parallel corpora in statistical machine translation systems. In *Proceedings of NODALIDA*, pages 331–341, Oslo, Norway, 2013a.
- Samuel Läubli, Mark Fishel, Manuela Weibel, and Martin Volk. Statistical machine translation for automobile marketing texts. In *Proceedings of MT Summit XIV*, Nice, France, 2013b.
- Mirko Plitt and François Masselot. A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague*

Bulletin of Mathematical Linguistics, 93:7–16, 2010.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL <http://www.R-project.org/>.

Sheila C.M. de Sousa, Wilker Aziz, and Lucia Specia. Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In *Proceedings of RANLP*, pages 97–103, Hissar, Bulgaria, 2011.

Martin Volk, Rico Sennrich, Christian Hardmeier, and Frida Tidström. Machine translation of TV subtitles for large scale production. In *Proceedings of EM+/CNGL*, pages 53–62, Denver, USA, 2010.

An Evaluation of Tools for Post-Editing Research: The Current Picture and Further Needs

Lucas Nunes Vieira
Newcastle University
School of Modern Languages, OLB
Newcastle upon Tyne, UK
l.nunes-vieira@ncl.ac.uk

Abstract

This paper presents a comparative evaluation of four tools that can be used to collect user activity data (UAD) in machine translation post-editing (PE) research: Tobii Studio, Translog-II, TransCenter, and PET. These tools are analysed here based on empirical data as a way of providing a picture of what the current state of research has to offer in terms of technology and investigation methods. After an analysis of the features offered by the tools, a summary is drawn and potential room for improvement in the field is identified.

1 Introduction

In view of the remarkable gains in quality achieved in Machine Translation (MT) in the past years, post-editing machine output is now growing to become an established translation modality in its own right and, as a result of this, an increasing number of researchers are starting to investigate the process of PE.

Research in the field tends to be particularly focused on the effort invested in the activity. However, investigating effort in PE, or in any other task, is a very challenging undertaking. Especially if temporal, technical, and cognitive effort (Krings, 2001) are taken into account, the use of tools that are able to log time, keyboarding, as well as potential indicators of cognitive effort (e.g. gaze data) becomes paramount in achieving research goals.

In order to cast light on the type of data obtained in PE investigations with the use of research tools currently available, four pieces of software are reviewed in the present paper: Tobii Studio (v.3.1.3), Translog-II (v.0.1.0.189),

TransCenter (v. 0.5), and PET (v. 2.0). These tools are chosen for analysis due to their prominence in previous research and their possibility of being exploited specifically for PE. All four tools are described in view of key- and time-logging features, and data visualisation aids, while eye-tracking features are only considered in view of Tobii Studio and Translog-II, since TransCenter and PET do not offer a built-in integration with eye trackers.

In describing these tools based on empirical data, the aim of this paper is to provide an overview of the current state-of-affairs in PE research technology and point to potential aspects that can be further improved in the field. Tools such as the ones used in Green et al. (2013) and Plitt and Masselot (2010), as well as the productivity-testing tool available in the context of the TAUS Dynamic Quality Framework¹ are not reviewed here, since, to the best knowledge of the author, they are not available to the general public. With regard to other tools that can be used for PE research, the CASMACAT (Ortiz-Martínez et al., 2012) and MateCat (Cattelan, 2012) workbenches were not included in the present review. Even though prototypes and beta releases of the tools are available, at the time of writing, the tools' development projects are yet to be finalised. Appraise (Federmann, 2012) and iOmegaT (Moran and Beregovaya, 2012), which are mainly focused on MT evaluation and PE productivity measurement, respectively, are two other tools that have not been reviewed. Even though these tools can be used for PE research, due to space and time limitations they could not be included in the analysis.

In the remainder of this paper, the criteria for analysis of the tools, the tasks conducted to investigate their usability, and a brief description

¹ <https://tauslabs.com/dynamic-quality>

of each tool are provided in Section 2. The tools are analysed in Section 3, and, in Section 4, suggestions are provided in terms of potential adaptations and areas where there is possible room for improvement in regard to technology that can be applied to PE research.

2 Context and Criteria for Analysis

In terms of the functionalities comprised in the tools that can be useful for research in PE, Tobii Studio and Translog-II are analysed based on the following eye-tracking-specific aspects:

- Amount of information comprised in gaze data logs and ease in computing measures at a sentence and/or sub-sentence level;
- Possibility of measuring gaze data quality.

As to features that do not necessarily involve eye tracking, all four tools are analysed based on the following criteria:

- Amount of information in time and key logs and ease in computing measures at a sentence level;
- Data visualisation aids;
- Customisation possibilities within the tools' environment.

The choice of these specific criteria is motivated by potentially challenging methodological aspects observed in previous research, such as computing per-segment PE time based on timestamps in the task video (O'Brien, 2011), and computing gaze data pertaining to ST and TT windows based on screen pixel positions (Hvelplund, 2011). Gathering UAD at the sentence level seems to be, overall, a common and yet challenging research need, frequently incurring in task designs where sample materials are exposed to subjects sentence by sentence, with no access to the whole text being granted – which is the case in Green et al. (2013) and Doherty et al. (2010), for example. Nevertheless, the criteria chosen in this review are by no means exhaustive, and the question of what exact set of features make for a good research tool in PE cannot be entirely solved in this paper.

The studies conducted to test the tools consisted of PE tasks with source text (ST) in Spanish and target text (TT) (MT output) in English. Spanish news texts of approximately 130 words each were translated into English with Google

Translate², and two professional translators post-edited the MT outputs. The eye-tracking equipment used with Tobii Studio and Translog-II is a Tobii X120 remote eye tracker.

2.1 Tobii Studio

Tobii Studio is the Windows-oriented eye-tracking software that accompanies Tobii eye trackers. Since the program does not have a built-in text editor, the screen-videoing mode needs to be used for PE tasks. When running in this mode, the program records everything that happens on the computer screen in the format of an .avi video, superimposing individuals' eye movements onto the recording. Data can be manipulated within the tool or exported in .tsv or .xlsx formats. Microsoft Word was the text editor used in combination with Tobii Studio. While this is arguably not the best text editor for PE research, this analysis only concerns features that apply specifically to Tobii Studio. In that way, Microsoft Word editing features and user interface (UI), as well as their usability for PE research, are beyond the scope of this paper.

2.2 Translog-II

Translog-II (Carl, 2012a) is a Windows-oriented software package designed specifically for translation process research (TPR). The package contains two tools: the *Supervisor*, and the *User*. Projects are set in the Supervisor, where any data produced can be visualised and manipulated. The User serves as the editing interface where participants carry out the task. Other than gaze data, the tool can also record keyboard and mouse events, as well as audio. In addition to the analysis possibilities presented within the environment of the tool, Translog-II data log files, which are saved in .xml format, can be further processed by a series of scripts included in the TPR database of the Centre for Research and Innovation in Translation and Technology (CRITT TPR-DB) (Carl, 2012b). Since these scripts are designed to process data in the format obtained with Translog-II, they are also taken into account in the present analysis, which is based on Version 1.2 of the scripts.

2.3 TransCenter

TransCenter (Denkowski and Lavie, 2012) is an open-source, web-based tool that allows different

² <http://translate.google.com/>

participants to carry out PE tasks remotely via a server. The tool logs time and keyboard/mouse activity at a sentence level. Subjective assessments of translation quality, difficulty, and usability can also be gathered through quality rating scales that are automatically included in the tool’s UI depending on the task chosen – if bilingual or monolingual PE, for example. Aggregate UAD for all participants, as well as data for each participant individually can then be accessed via report files generated by the tool in both .csv and .html formats. Since the tool is web-based, TransCenter can be accessed on any platform.

2.4 PET

Out of the four tools considered, PET (Post-Editing Tool) (Aziz et al., 2012) is the only one designed specifically for PE. Similarly to TransCenter, PET is open-source and platform-independent. In addition to recording time and effort indicators at a segment level, PET also allows users to perform assessment tasks based on configurable rating scales and criteria. UAD generated with PET is saved in .xml format.

3 Analysis

In the following section (3.1), Tobii Studio and Translog-II are analysed in view of eye-tracking-specific features. Since TransCenter and PET do not log gaze data, these tools are not analysed in this section. In sections 3.2 and 3.3, all tools are taken into account.

3.1 Gaze data

3.1.1 Amount and type of information in gaze data logs

In Tobii Studio’s data log file, gaze events are classified as ‘fixation’, ‘saccade’ or ‘unclassified’, and each event is accompanied by information such as the positions of both right and left eyes on screen, left and right pupil sizes, distance of both eyes from the screen, as well as the gaze event’s duration in milliseconds. Clusters of gaze events that are identified as a single fixation or saccade receive a respective index number. Gaze events are grouped into fixations based on fixation-filter settings that are configured by the user. An extract of the data log file generated with Tobii Studio is presented in Table 1.

Recording Timestamp	FixationIndex	SaccadeIndex	GazeEventType	GazeEventDuration
430	5		Fixation	158
439		5	Saccade	42

Table 1. Extract of Tobii Studio data log file

As for Translog-II, the .xml results file with UAD contains information such as source and target (MT output) texts, the task (if ‘translating’ or ‘post-editing’, e.g.) as well as keyboard, mouse and time logs, cursor positions, and gaze data. In terms of gaze data, the file includes information such as the timestamp associated with each gaze event, positions of right and left eye on screen, as well as pupil size.

With respect to differences between Tobii Studio and Translog-II in terms of the type of gaze data generated, the latter – being a tool specifically designed for TPR – automatically records information pertaining to the particular window (ST or TT) a given gaze event is related to. In Translog-II, gaze events can be filtered into fixations based on the CRITT TPR-DB scripts. After aligning ST and TT with the jdtag³ tool, these scripts can be used to produce, among other things, a series of unit tables with process and product data as well as files that can be used in external tools for part-of-speech (POS) tagging and syntactic parsing. Below is an example of a fixation data (FD) table generated with the CRITT TPR-DB scripts.

FIXid	Time	Dur	Win	Cursor	ParaK	Edit	EditID	STid	TTid
29	7987	1066	1	878	0	---	---	159	157
30	10389	142	2	148	142	—	10+	28	29
31	10561	366	2	55	94	of	11+	12	12

Table 2. Translog-II Fixation Data

As shown in Table 2, similarly to Tobii Studio, each eye fixation in Translog-II has an individual ID and is accompanied by duration and timestamp (the columns ‘Dur’ and ‘Time’, respectively). In Translog-II, however, thanks to the gaze mapping functionality of the tool, it is also possible to know what word in the text the fixation refers to. Each word in the text is given an ID number, and ST and TT word ID pairs (columns ‘STid’ and ‘TTid’) associated with fixations are also displayed in the table.

Still in regard to gaze data, also generated by the CRITT TPR-DB scripts is a table with fixation units (FU). A concept proposed by Carl and Kay (2011), FUs are clusters of fixations that, together, represent one meaningful sequence. FU

³ <http://code.google.com/p/jdtag/>

tables have information on the time each unit started, its duration, as well as the amount of time for which reading and typing took place in parallel.

Data exported from both Tobii Studio and Translog-II could arguably be deemed to be in interoperable formats overall, since the former exports data in .tsv format, and the latter in .xml. As regards the replay function, however, the fact that tasks can be replayed based on the .xml file in Translog-II arguably allows for an easier storage and transport of data. In Tobii Studio, by contrast, tasks are replayed from .avi files, which tend to be considerably large and hence potentially difficult to store and transport.

3.1.2 Computing gaze-data measures at a sentence and/or sub-sentence level

As to computing measures for specific moments of the task, Tobii Studio allows the possibility of selecting video passages and marking them as ‘scenes’. Statistics referring to specific ‘areas of interest’ (AOIs) on the screen (ST and TT windows, say) within a scene are then computed. In that way, if specific sentences or phrases can be identified in the text as AOIs, it is possible to draw a polygon around the corresponding area and obtain data pertaining only to the particular area selected.

In Translog-II, by contrast, each fixation is automatically mapped to specific ST and TT words with the use of the CRITT TPR-DB scripts. Even though the quality of the gaze mapping might also depend on the precision of the eye tracker used, this functionality arguably allows for an easier consideration of gaze data at a sentence and/or sub-sentence level. In addition, Translog-II offers the possibility of correcting gaze mapping manually after conducting a task – a functionality not offered by Tobii Studio.

3.1.3 Measuring gaze data quality

In experimental designs where eye-tracking data is used, an important step in the analysis process is to account for data quality. In this respect, Tobii Studio has a built-in measure that assesses the confidence that a given gaze event is in fact valid, generating values that can range from 0 (high confidence) to 4 (no eye found).

Measures of data quality can also be computed based on information in Tobii Studio’s data log file. In Hvelplund (2011:103-107) e.g., where a previous version of Tobii’s eye tracking soft-

ware was used, mean fixation duration, gaze-sample-to-fixation percentage, and a ratio between gaze time on screen and total production time have been used as indicators of gaze data quality. In Translog-II, a ratio of gaze events happening in windows 1 and 2, and gaze events that did not happen in any window, i.e. events that have 0 as a window value, constitutes another potentially interesting strategy to measure data quality informally suggested to the author by Translog-II developers.

Overall, with respect to gaze data, while Tobii Studio and Translog-II generate raw output files with similar information, the scripts in the CRITT TPR-DB database allow for a number of further automatic data analysis stages which, in Tobii Studio, would arguably involve lengthy processing steps.

3.2 Key and time logs

3.2.1 Amount and type of information in key and time logs

In addition to gaze data, Tobii Studio also logs keyboarding and mouse clicks, which can be found together with gaze data in the same log file. All these events are associated with their respective timestamp based on the task video.

With respect to Translog-II, CRITT TPR-DB unit tables include a keystroke-data (KD) table, as well as production-unit (PU) and alignment-unit (AU) tables. The KD table includes information on the number and type of editing operations performed (insertions, deletions) and the words in the ST and post-edited text associated with them. Similarly to the concept of FU, PU are clusters of editing operations that can be regarded as a single unit. AU tables, in turn, contain process and product data pertaining to aligned source and post-edited units, i.e. the edits performed and the aligned result of these edits.

With respect to TransCenter, measures such as edit, keypress and mouseclick counts are recorded per sentence. The tool also records editing time, and how each sentence is scored by participants based on 1-5 scales.

As regards PET, the tool distinguishes between white-space, non-white-space and control keyboard events, classifying each event according to a fine-grained list of categories, including e.g. ‘navigation-keys’ and ‘paste-keys’. In addition, it offers a few functionalities that are different to the ones found in other tools, such as au-

tomatically labelling clusters of insertions and deletions as ‘substitutions’ and ‘shifts’, and computing Human Translation Edit Rate (HTER) (Snover et al., 2006) as a built-in effort indicator. PET also logs sentence/segment-specific measures of editing time.

3.2.2 Computing key and time measures at a sentence level

In terms of time-logging, Tobii Studio simply offers the timestamp associated with each event recorded by the tool. One way of computing measures of time at a sentence level with Tobii Studio is by considering the timestamps in the task video associated with the moments when participants began and finished editing each sentence. In this respect, if the task is not carried out on a sentence-by-sentence nature where each segment/sentence needs to be confirmed before moving on to the next, computing sentence-specific time measures in Tobii Studio constitutes an arguably unreliable approach, since it would be hard to collect such measures without distinct time delimitations between sentences.

With respect to Translog-II data, due to the tool’s gaze mapping functionality, information on time can be obtained for each FU or PU, for example. In addition, when setting up an experiment in Translog-II, the ST can be divided into translation units that are displayed separately according to settings established by the researcher, such as a time limit for which the segments will be displayed. The time spent on each unit can then be observed in the data log file that is generated after a task is completed. However, in the context of this evaluation, this functionality did not seem possible to be used for PE, since only the ST seems to be breakable into units, and not both ST and TT (MT output). In this respect, PET and TransCenter seem to be the only tools analysed that offer an automatically computed measure of time per ST-TT segment, which can be useful in PE task designs where sentences are established as units for analysis.

In sum, while time measures at a sentence level need to be computed based on timestamps in Tobii Studio, in Translog-II these measures can be computed for ST-based units, or for sentence and sub-sentence units based on fixations and/or keyboard events. PET and TransCenter, in turn, offer automatically computed key and time measures per ST-TT segment.

3.3 Data visualization aids

In addition to quantitative data that can be exported from Tobii Studio for each task, the software has a number of different graphic representations of data that can be explored. Tables and charts can be generated and gaze events can be viewed in the form of gaze plots, where eye fixations can be observed on a still screen capture extracted for a given timespan in the task video.

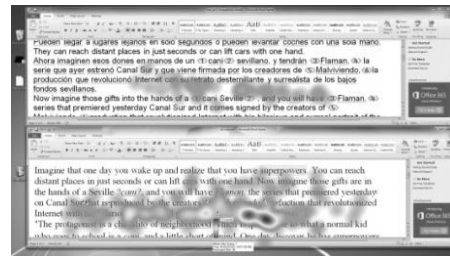


Figure 1. Tobii Studio Heat Map

Another visualisation option offered by Tobii Studio are heat maps (Fig. 1), where a colour representation – ranging from green (cool) to red (hot) – indicates the areas of the screen that received more gaze events.

One of the most prominent visualisation options in Translog-II is what is referred to as the ‘linear view’ (Fig. 2), where the editing process can be observed linearly with different editing events (including eye fixations), represented by different symbols and colours.

```
{2:Margare}{2:et That}{2:at cher,}{1:vera de}{1:et That}{1:Margare}{1:primera}{1:
mujer}{1:e alcan}{1:ó el ca}{1:cargo}{1:e prime}{1:minist}{1:
primer}{1:ministr}{1:Reino U}{1:rimer m}{2:tcher,}{2:he firs}{2:woman}{2:to
rea}{2:reach}{2:he prim}{2:me mini}{2:ch the}{2:reach}{2:e prime}{1:as.
}{2:r of Un}{1:as.
}[▼][▲]•post{2:f 1979 •}{2:of prio}{2:minist}{1:as.
}{1:as.
}{2:That ch}{2:Kingdom}{2:and Bri}{2:ritish}{2:nd Brit}{2:olitica}{2:life tr}{1:
Según}{1:Margare}{1:y que t}{1:Reino}{1:tro de}{1:ministr}{1:e
prime}{1:Margare}{1:Margare}{1:Margare}{1:Margare}{1:as.
}{1:as.
}[▼][▲]◀◀in{2:ctim of+}{2:r in t t}{1:as.
```

Figure 2. Translog-II Linear View

In the linear view extract in Fig. 2, keyboard, mouse and fixation events are displayed. Portions of fixated text are displayed inside brackets, where the number before the colon represents the window where the fixation occurred – 1 (one) refers to the ST, and 2 (two) refers to the TT. Two consecutive triangles pointing downwards and upwards represent a click. With regard to keyboard events, dots represent spaces, triangles pointing backwards represent deletions, and insertions are displayed simply as the letters that were actually typed by the participant.

Another way of viewing data in Translog-II is by replaying the task via the .xml log file. Data

can also be viewed in the format of a pause plot, where keyboard pauses can be observed in a graph. A screen capture of the replay function in Translog-II is presented in Fig. 3, where the focus of gaze data is represented by a circle and its mapping by a rectangle over the respective portion of text being fixated.

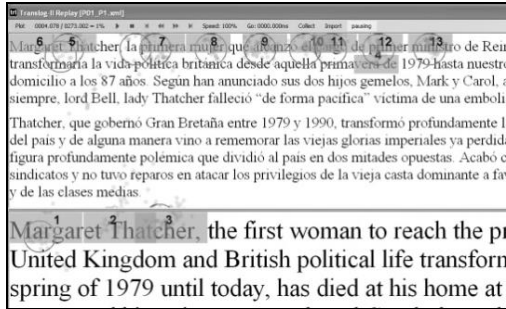


Figure 3. Translog-II Replay Function

Translation progression graphs (TPGs) (see Fig. 5) present another possibility of visualising Translog-II data. A feature exclusive of Translog-II, these graphs can be generated with the statistical package R⁴ based on the tables created with the CRITT TPR-DB scripts.

TPGs can be very informative in denoting combined reading and production patterns. Perhaps to make such graphs more useful in the context of PE, adding reference to the post-edited text (and not only the ST) would be desirable.

With respect to TransCenter, the tool enables a sequential edit-by-edit visualisation of the PE process through ‘edit trace reports’ (Fig.4). The tool also displays aligned ST, MT output and post-edited sentences together with sentence-specific UAD.

Time	Sentence 4 Edits
Initial	How does Mary Shelley finish off?
1370010899990	How does Mary Shelley finish off?
1370010899990	How w Mary Shelley finish off?
1370010900057	How wo Mary Shelley finish off?
1370010900108	How wou Mary Shelley finish off?
1370010900252	How woul Mary Shelley finish off?
1370010900348	How would Mary Shelley finish off?

Figure 4. TransCenter Edit Trace Report

In regard to PET, no pre-set data visualisation options seem to be available within the environment of the tool. In this respect, while TransCenter has interesting visualisation possibilities not offered by PET, the latter seems to provide more detailed keyboard data, which can always be explored by the researcher in external data-analysis tools.

Overall, in terms of visualisation possibilities, heat maps figure as a distinctive feature of Tobii Studio, while TPGs constitute a feature that can be especially useful for PE research and which is offered exclusively by the CRITT TPR-DB scripts. A linear view of the editing process is offered by Translog-II, with a similar and less detailed alternative being offered by TransCenter in the form of edit trace reports.

3.4 Customisation Possibilities

In this section, customisation options presented within the environment of the tools are analysed. While PET and TransCenter are both open-source tools, the analysis presented here focuses on settings that can be customized without recourse to the tools’ source code.

Since, in the context of PE tasks, Tobii Studio needs to be used with an external text editor, the customising possibilities presented by the tool itself are limited to fixation-filter settings and data visualisation options.

In addition to data visualisation options, such as the colour representation and choice of events to be included in the linear view, Translog-II presents a few task-related customisation possibilities, such as choosing reading, translating or writing as linguistic tasks, and having the window panes displayed accordingly. In the replay mode in Translog-II, it is also possible to choose the FixMap option, where gaze mapping can be manually corrected. With respect to the data log file generated with Translog-II and how it can be processed, a number of possibilities are available to the researcher by manipulating the CRITT TPR-DB scripts, including the configuration of fixation-filter settings, which can be recomputed with the command ‘remap’.

Being a tool designed specifically for PE, PET presents a number of potentially useful options that can be explored in the specific context of PE research, such as displaying buttons that allow participants to either accept the MT output as is or discard it altogether – actions that can be tracked later in the results log file generated by the tool. PET also has a drag-and-drop functionality that allows text to be moved both within an active unit, as well as from any segment in the text into the active TT unit being edited.

In comparison with PET, TransCenter seems to offer fewer customisation possibilities that can be configured with no recourse to the tool’s source code. No instructions were found on how

⁴ <http://www.r-project.org/>

to change rating scales or the way panes are displayed in the editing interface, for example. On the other hand, TransCenter is the only tool out of the ones analysed that can be accessed via a server. While PET can also be used remotely by participants, being able to access TransCenter with a username and password on a web browser arguably facilitates the data-collection process, which can be controlled remotely by the project's administrator.

PET also allows access to dictionaries and other reference material within the environment of the tool. While this functionality could also be observed for Translog-II in the tool's documentation, this feature did not seem to be included in the version of Translog-II analysed, nor in a subsequent version (v.0.1.0.191) released after the experiments reported in this paper had been conducted. This renders PET the only tool reviewed to have a functional integration with reference materials.

4 Conclusion and Further Issues

A summary of the functionalities observed for each of the tools described can be observed in Table 4.

As can be seen from the descriptions provided, both Tobii Studio and Translog-II allow an analysis of gaze and keyboard/mouse data both quantitatively, with the generation of tables and statistics, and qualitatively, with features such as the replay function, the linear view and TPGs. PET figures as a powerful option mainly for quantitative investigations specifically on PE, presenting pre-set configurable functionalities that are particularly useful for gathering human assessments as well as measuring temporal and technical effort in PE, which, as with TransCenter, can be considered at a sentence/segment level. With regard to TransCenter, one of the main differentials of the tool seems to be the fact that it is web-based, which allows for an arguably easier running of research tasks.

Tobii Studio constitutes an option that can be adopted when research experiments need to be more ecologically valid and not necessarily strictly controlled, since any commercial CAT tool, such as Trados⁵ or memoQ⁶, can be used in combination with Tobii Studio. In this respect, combining the use of PET or TransCenter with

Tobii Studio also figures as a potentially interesting possibility in which all the PE-specific UI functionalities of PET and TransCenter can be exploited in eye tracking studies.

As regards file formats, all four tools seem to meet good levels of interoperability, with data being saved in formats such as .csv, .tsv, and .xml.

Features	Tobii Studio	Translog-II	TransCenter	PET
Pupil size recording	x	x	n/a	n/a
Built-in gaze data validity measure	x		n/a	n/a
TPGs		x	n/a	n/a
Heat maps	x		n/a	n/a
Gaze plots	x	x	n/a	n/a
Quant. info on saccades	x		n/a	n/a
Automatic gaze mapping		x	n/a	n/a
Manual gaze mapping		x	n/a	n/a
Pause plots		x		
Linear view		x	x	
Replay function	x	x		
Integr. w/ audio recorder**	x	x		
Integr. w/ reference materials		x*		x
Time recording per segment		x	x	x
Human assessments			x	x
Inference on shifts and subs. (keyb.)				x
Drag and drop				x
Division of text into PE units			x	x
Locking/hiding PE segments				x
Server			x	
Platform-independent			x	x
Open-source			x	x

*Not observed in v. 0.1.0.189

**Not used in the context of this paper

Table 4. Summary of Features

In terms of qualitative analyses, it seems that Translog-II is able to provide a larger number of possibilities to be exploited. In the context of PE, TPGs would arguably be more informative if the post-edited text is also displayed. For demonstration purposes, an adapted version of such graphs is presented in Fig. 5 together with retrospective verbalisations. In these graphs, the y-axis shows ST words in sequence, dark circles represent fixations in the ST, lozenges represent fixations in the TT, black characters represent insertions and red ones represent deletions.

When accompanied by spoken data (in this case, retrospective think-aloud protocols) and the post-edited text, TPGs potentially allow for a powerful and in-depth analysis of the PE process. In this example, it is possible to observe, for instance, that in the time interval shown, the participant had few fixations on the ST relating to the text passage displayed, as signalled by the small

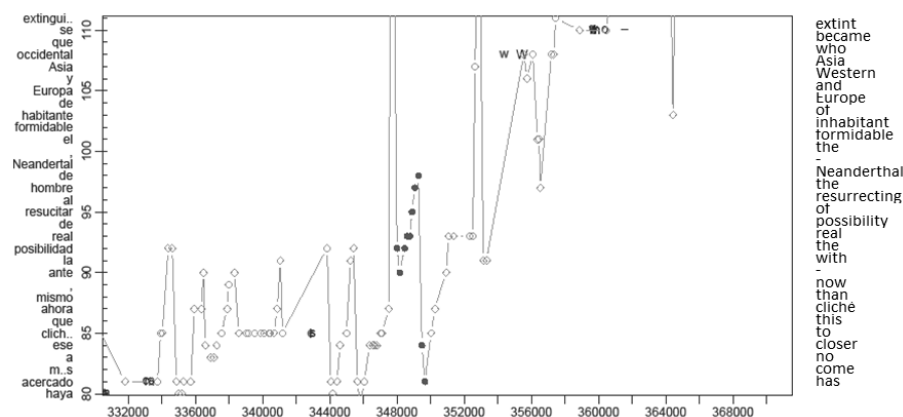
⁵ <http://www.trados.com/en/>

⁶ <http://kilgray.com/products/memoq>

number of dark circles within the range of the graph. It is also possible to observe that the process of editing this passage was far from linear, which is demonstrated by the saccades in the map, where the participant seems to be reading backwards and forwards in an overlapping fash-

despite its potential utility, renders debatable the reliability of this measure.

As a relatively new activity, research methods currently available for PE seem to heavily draw on more established areas such as reading and traditional translation. In view of this, it seems



Not closer to the cliché, I'm not quite sure, no closer to the cliché. I was trying to make sense out of it, that's really why I've added come. This cliché, I'm making explicit, perhaps not necessary, but the fact that the Hollywood thing is a cliché. Just capitalising Western Asia, I think it's better.

Figure 5. Translog-II translation progression graph with TT and spoken data

ion.

By referring to retrospective spoken data pertaining to the same text passage covered in the graph, it is possible not only to provide a clearer indication of the changes taking place – since deletions and insertions frequently overlap in the graph, hindering full comprehension – but also show the possible mechanisms behind the edits performed.

In terms of other features that could be implemented in tools that can be used for PE research, computing the amount of mouse hovering events and mapping them to their corresponding words in the text figures as a potentially interesting function to be explored. This approach has been used for PE by Green et al. (2013), who mention previous studies where mouse hovering has been shown to correlate with eye-tracking data. In view of the constraints imposed by eye tracking due to the need for specialised equipment and appropriate conditions, it would perhaps be interesting to see automatically computed measures of mouse hovering in freely available research tools that can be used for PE. It is noteworthy, however, that studies looking at the correlation of gaze data with mouse hovering specifically for PE are apparently lacking, which,

that in-depth studies into the operational underpinnings of PE would lead to better strategies of data collection that reflect the PE activity more directly. The amount of crossing between ST and MT output is an example of a potential measure of effort in PE that is arguably under-explored. In addition, it seems that only recently there have been initiatives at developing data-collection tools that are able to mimic more advanced CAT

functionalities offered in commercial CAT software, which is an aspect that the CSMACAT and MateCat projects aim to attend to. In this respect, the controlled lab conditions enabled by research tools such as the ones reviewed in this paper might hinder more valid investigations into effort, since, when using these tools, participants are not able to count on functionalities that they would normally be able to use in real-world contexts, such as interactive editing features and on-the-fly quality assurance checkers, for example.

As future work, it would be interesting to expand this review by including the analysis of other tools. Data obtained with other studies could also be considered in order to check to see if research needs are met across a wider and more diverse context.

Acknowledgments

This evaluation is part of a project supported by the School of Modern Languages at Newcastle University. Particular gratitude is extended to Dr Francis Jones, Dr Michael Jin, and Dr Ya-Yun Chen for their collaboration.

References

- Aziz, W., Sousa, S. C. M., and Specia, L. 2012. PET: a tool for post-editing and assessing machine translation. *The Eighth International Conference on Language Resources and Evaluation, LREC'12*, Istanbul, Turkey. May 2012, 3982-3987.
- Carl, M. 2012a. Translog-II: a program for recording user activity data for empirical reading and writing research. *The Eight International Conference on Language Resources and Evaluation, European Language Resources Association*, Istanbul, Turkey. May 2012, 4108-4112.
- Carl, M. 2012b. The CRITT TPR-DB 1.0: A Database for Empirical Human Translation Process Research. *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*. ed. / Sharon O'Brien; Michel Simard; Lucia Specia. Stroudsburg, PA : Association for Machine Translation in the Americas (AMTA), 2012. 9-18.
- Carl, M. and Kay, M. 2011. Gazing and Typing Activities during Translation: A Comparative Study of Translation Units of Professional and Student Translators, *Meta: Journal des traducteurs/Meta: Translators' Journal*, 564, 952-975.
- Cattelan, A. 2012. MateCat. *D4.1 First Version of MateCat Tool*. Available at http://www.matecat.com/wp-content/uploads/2013/01/MateCat-D4.1-V1.1_final.pdf [Accessed 16 July 2013].
- Denkowski, M. and Lavie, A. 2012. TransCenter: Web-Based Translation Research Suite, *AMTA 2012 Workshop on Post-Editing Technology and Practice Demo Session*, 2012.
- Doherty, S., O'Brien, S., and Carl, M. 2010. Eye tracking as an MT evaluation technique. *Machine translation*, 24(1), 1-13.
- Federmann, C. 2012. Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations. *The Prague Bulletin of Mathematical Linguistics (PBML)* 98, 25-35.
- Green, S., Heer, J. and Manning, C.D. 2013. The Efficacy of Human Post-Editing for Language Translation. *ACM Human Factors in Computing Systems (CHI)*, 2013.
- Hvelplund, K.T. 2011. *Allocation of Cognitive Resources in Translation: an Eye-Tracking and Key-Logging Study*. Ph.d.-afhandling. Copenhagen Business School Copenhagen
- Krings, H. P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Vol. 5. Kent, Ohio: Kent State University Press.
- Moran, J. and Beregovaya, O. 2012. iOmegaT – an adapted open--source CAT tool to measure. MT post--edi ng produc tivity in enterprise deployments. Demo Poster in: *AMTA 2012*. Available at http://m25s17.vlinux.de/098709809/AMTA2012_DemoPoster.pdf [Accessed 16 July 2013].
- O'Brien, S. 2011. Towards predicting post-editing productivity. *Machine translation*, 25(3), 197-215.
- Plitt, M. and Masselot, F. 2010. A productivity test of statistical machine translation post-editing in a typical localization context. *Prague Bull Math Linguist*, 93, 7-16.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *The 7th Conference of the Association for Machine Translation in the Americas*, 223-231.
- Ortiz-Martínez, D., Sanchís, G., Casacuberta, F., Alabau, V., Vidal, E., Benedí, J. M., González-Rubio, J., Sanchís, A. and González, J. 2012. The CASMACAT Project: The Next Generation Translator's Workbench. *The 7th Jornadas en Tecnología del Habla and the 3rd Iberian SLTech Workshop (IberSPEECH)*, 326-334.

Online production of HQ parallel corpora and permanent task-based evaluation of multiple MT systems: both can be obtained through iMAGs with no added cost

Lingxiao Wang Christian Boitet

UJF, LIG-GETALP, BP 53

41 rue des Mathématiques, Domaine Universitaire

38041 Grenoble Cedex 9

Lingxiao.Wang@imag.fr, Christian.Boitet@imag.fr

Abstract

An interactive Multilingual Access Gateway (iMAG) dedicated to a website S (iMAG- S) is a good tool to make S accessible in many languages immediately and without editorial responsibility. Visitors of S as well as paid or unpaid post-editors and moderators contribute to the continuous and incremental improvement of the most important textual segments, and eventually of all. In this approach, pre-translations are produced by one or more free machine translation (MT) systems. Continuous use since 2008 on many websites and for several access languages shows that a quality comparable to that of a first draft by junior professional translators is obtained in about 40% of the (human) time, sometimes less. There are two interesting side effects obtainable without any added cost: iMAGs can be used to produce high-quality parallel corpora, and to set up a permanent task-based evaluation of one or more MT systems.

1 Introduction

An iMAG is an interactive Multilingual Access Gateway very much like Google Translate at first sight: one gives it a URL and an access language and then navigates in that access language. When the cursor hovers over a segment, a palette shows the source segment and proposes to contribute by correcting the target segment, in effect post-editing a MT result or improving on a previous post-edition. With Google Translate, the page does not change after contribution, and if another page contains the same segment, its translation is still the rough MT result, not the polished post-

edited version. The more recent Google Translation Toolkit enables one to MT-translate and then post-edit online full web pages from sites such as Wikipedia, but again the corrected segments don't appear when one later browses the same page in the access language.

By contrast, an iMAG- S is *dedicated* to an *elected website* S , or rather to the *sublanguage* empirically defined by the textual content of one or more URLs constituting S . The iMAG- S contains a translation memory (TM) and if possible a specific, pre-terminological dictionary (pTD) (Daoud et al., 2009), both dedicated to the elected sublanguage. Segments are pre-translated not by a unique MT system, but by a (selectable) set of MT systems. Systran, Reverso and Google Translate have been mainly used as well as Neon for Chinese-English, but specialized systems developed from the post-edited part of the TM, and based on Moses (Koehn et al., 2007), are also used in our gateway.

The online contributive platforms SECTra_w (Huynh et al., 2008) and PIVAX (Nguyen et al., 2007) are used to support the TMs and pTDs. Translated pages are built with the best segment translations available so far. While reading a translated page, it is possible not only to directly post-edit the segment under the cursor, but also to seamlessly switch to SECTra_w online post-editing (PE) environment, equipped with filtering and search-and-replace functions, and then to go back to the reading context. To illustrate our points, we will use an iMAG created for the website of our lab (400 researchers, 25 teams).

Since 2008, we have regularly added iMAGs to our platform, and found that two interesting side effects are obtainable without any added cost: iMAGs can be used to produce high-quality parallel corpora, and to set up a permanent task-based evaluation of one or more MT systems.



Figure 1: Access in Chinese of to the LIG lab website

2 Typical scenario of use

2.1 Multilingual access to a website

Figure 1 shows the iMAG access interface to the LIG lab website. We choose Chinese as the access language from the pull-down menu. One or more free MT servers, in this case Google Translate and Systran, produce initial translations.

2.2 Post-editing and scoring on the page

As shown in Figure 2, when the mouse pointer hovers on a segment (title, sentence, menu item), an interactive palette pops up. It's dialogue box displays the source language content (in blue), and users can post-edit and evaluate the text in the access language.

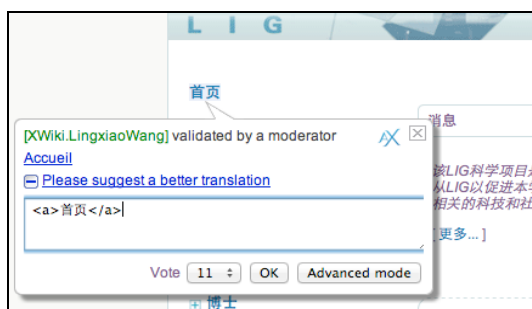


Figure 2. Direct PE of translation results

All visitors of the page can contribute by post-editing. However, only registered users can use the "Advanced mode". If the TM contains several post-editions for one segment, the system selects which to use in the translation page, based on highest score and then on most recent time.

2.3 Visualization of translation reliability

In the translated page, users can see the estimated *reliability* of each segment by checking the "Reliability" checkbox. As shown in Figure 3, colored brackets { _..._ } enclose each translated segment. If a user post-edits this page, colors¹ change based on the user's profile². If one clicks the "Original" button (in the upper right corner of figure 3), the left side of the browser window displays the page in the access language, and the right side the original page.

2.4 Post-editing TMs in "Advanced Mode"

The "Advanced Mode" offers a translation editor interface similar to those of translation aids and commercial MT systems, that makes post-editing much faster than in the presentation context. Not yet post-edited segments can be selected, and global search-and-replace is available. Figure 4 shows a screenshot of SECTra_w PE interface.

When an iMAG-S is created, we select several MT systems for proposing *pretranslations*, and set the *preferred* one. That can be changed later. From the post-editing interface, it is possible to perform various operations on the TM:

- MT results: discard a MT result, call again one of the selectable MT systems, and use an MT result as current post-edition³.
- Post-editions: discard a post-edition, use one as preferred in the current context.

¹ Green: privileged users; Orange: anonymous users; Red: MT output (the translation results have never been edited).

² For that, the page must be refreshed.

³ That result is then moved to the PE cell.

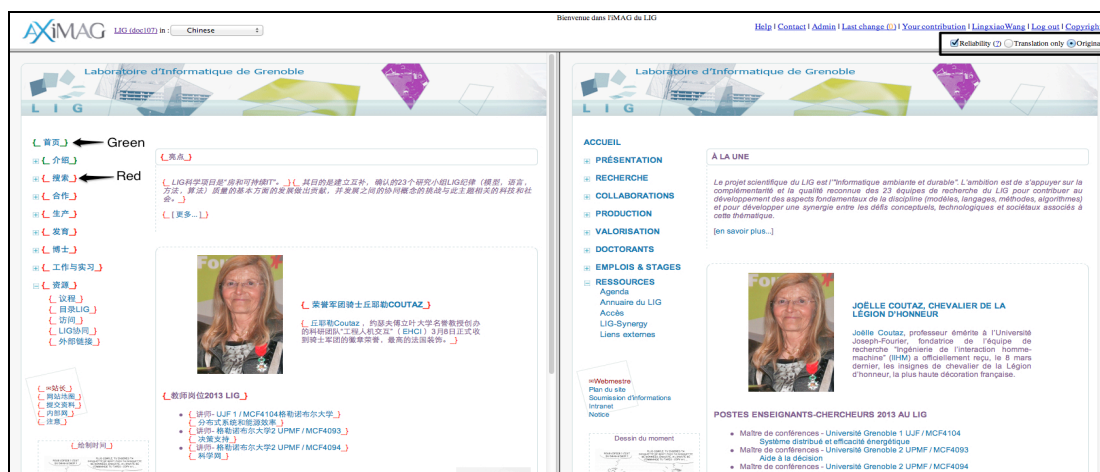


Figure 3. iMAG page display in “Reliability” + “Original” mode

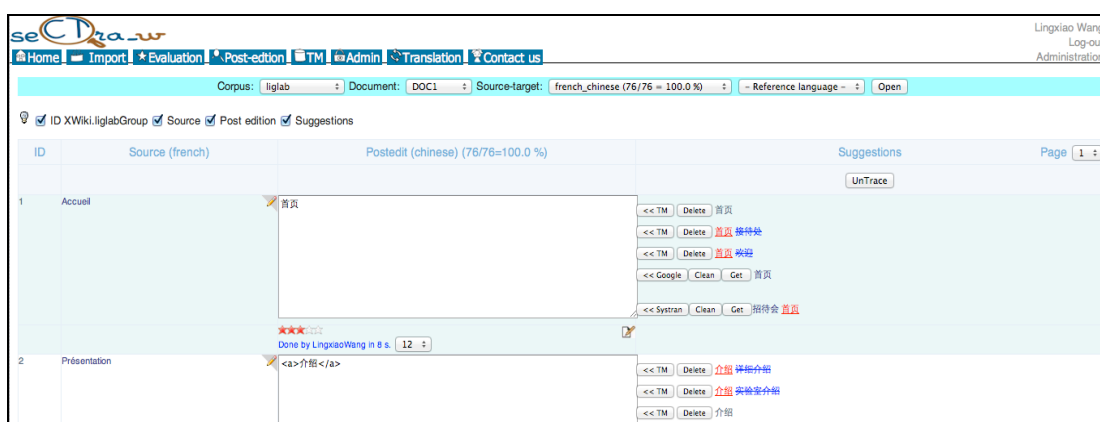


Figure 4. Advanced mode (SECTra_w screenshot)

As shown in Figure 4, users can visualize and compare the edit distances between the chosen current translation and the MT and PE results for that segment, contained in the translation memory, using the "Trace/Untrace"⁴ button.

2.5 User profiles, moderation, scoring

The admin assigns a profile to each registered user: *reliability* (bilingual, professional translator, translator certified for site S) and *quality score* for each language pair (from 0 to 20). That is based on language tests (A1-...C2, or ILTS, TOEFL, TOEIC, etc.), or on professional levels. *Moderators* are contributors competent enough in the domain of S and ‘blessed’ by S.

The *quality score* of a post-edited segment is, by default, that of its post-editor. It may be changed by the post-editor herself (self-evaluation), and also later by an admin or by a moderator.⁵

⁴ A "mixed character/word edit distance" is used.

⁵ There is also a subjective evaluation environment, where several judges can participate, assigning classical scores.

3 Conclusions after 4 years of use

After the first four years of use, there are over 80 demo iMAGs in operation, sharing 3 TMs, and 6 dedicated iMAGs, with their own TMs. There are 8 source languages, and websites can be accessed in more than 10 languages with post-editing support. There are more than 820,000 segments, and about 45% (370,000 +) segments have been post-edited by contributors. Most parallel segments are English-French, English-Chinese, and Chinese-French. We give some statistics in Table 1.

3.1 Reliability and quality indicators

It is quite difficult in practice to maintain a website in more than one language. Take for example the website of Figure 1, which an admin tries to maintain also in English: hitting the “English” button directs to web pages that are still 85% in French, while the 15% portion in English is far from perfect. By contrast, using the iMAG button and choosing English shows pages 100% in Eng-

lish, and the reliability can be shown for each segment. The reliability and score of each segment are also visible in advanced mode. More important, the overall quality of the post-edited segments (green or orange special brackets if

shown), estimated by teachers of English and bilinguals, is at least as good as that of translations (when available) found in the static “English version”.

Language pair (L1→L2)	Bi-segments	Source Words L1 (Standard p.)	Target Words L2 (Standard p.)	Size L1	Size L2
English → French	121 074	2 542 731 (10 170 p.)	2 613 351 (10 453 p.)	10,1MB	10,4MB
English → Chinese	208 106	4 370 530 (17 482 p.)	6 063 942 (151 159 p.)	19,1MB	17,6MB
French → English	29 079	627 661 (2510 p.)	610 098 (2 440 p.)	4MB	3,9MB
French → Chinese	10 890	228 703 (914 p.)	317 322 (793 p.)	1,5MB	1,25MB
Chinese → English	2 013	58 656 (146 p.)	42 275 (169 p.)	240KB	263KB
Chinese → French	10 062	291 192 (727 p.)	211 185 (844 p.)	874KB	1MB

Table 1. Parallel segments obtained (we count the number of Chinese characters for Chinese) (250 words/page in English/French, or 400 Chinese characters/page.)

The “trick” behind this is simple: no target pages are kept anywhere in our system. Only individual segments are kept (in the dedicated TM). Pages in target languages are dynamically built using the best translation (MT output or PE) available for each segment. We decide which string is the best for a segment based on the reliability of the post-editor⁶ (3 stars: amateur translator, 4 stars: professional translator, 5 stars: certified translator), and on the quality score (from 0 to 20) of MT pre-translations and post-editions.

3.2 Gains in human time (usage value)

From the point of view of the human time spent, how efficient is this method? As the first author is Chinese, he has experimented with French→Chinese and with Chinese→English on segments of a shared TM called Demo2 (including some French and Chinese short articles). We give in Table 2 the statistics gathered during one week (21-27 January 2013). During this week, 1853 segments were post-edited from French into Chinese, and 625 segments were post-edited from Chinese into English, and then an amateur translator translated the same segments without the help of iMAG. We recorded the time taken in each case, and compared the results.

It is well known that one should always post-edit into one’s native language: quality should be better and time shorter. However, the measures above seem not to confirm the second point. About 1342/1757=76% of the time is saved in the Fr→Zh direction, and 312/1464=78,6% in

the Zh→En direction. But close inspection of the results reveals that, as expected, the Zh post-editions are quite good⁷, while the En post-editions are not always exact and very often ungrammatical. Another step of revision by a native English speaker would be necessary before attaining the same translation quality as for Fr-Zh.

Note that these gains are in agreement with early experiments done in 2005 by Jeff Allen⁸ with professional translators post-editing into their native language Systran outputs.

We would like to speak here of *usage quality*, or even better of *usage value*. Since the early days where MT was deployed (Hutchins and Somers, 1992), it has been noted that linguistic quality and usage value do not correlate with each other. In fact, *while the linguistic quality of MT outputs is often judged to be very low by linguists and translators, their usage value is often quite high*.

With our setting, the linguistic quality of the post-edited segments (if post-editors work into their native language) is comparable with that of segments translated by junior professional translators having no special knowledge of the “sub-language” of the accessed website⁹. An interesting remark is that, whatever the PE direction, people seem to have some internal sense of “expected speed”, that ranges between 15 minutes per page to 25 minutes for the most scrupulous.

⁷ Five Chinese students to help us verify the results, they proved the correctness and readability.

⁸ See <http://www.oocities.org/mtpostediting/>

⁹ The second author has worked as technical translator and revisor and is in a position to make that kind of judgement.

⁶ 1 star: word for word translation, 2 stars: result of MT.

Language pair	Human PE time	Human first draft time	Segments	Source words (Standard pages)	Target words (Standard pages)
French→Chinese	415 mins	1757 mins	1 853	38 913 (155 p.)	46 648 (116 p.)
Chinese→English	312 mins	1464 mins	625	12 853 (32 p.)	8 568 (34 p.)

Table 2. Statistics from 1-week experiment (we count the number of Chinese characters for Chinese)

No	Pseudo Doc	Source	Cible	Stars	Notes
1	DOC16	la salle du haut conseil située au 9ème étage de l'institut du monde arabe, dans le 5ème.	位于巴黎5区的阿拉伯世界博物馆10楼高级理事会议厅。	3	20
2	DOC16	le 24ème étage de la tour zamansky, l'université pierre et marie curie, sur le campus de jussieu, dans le 5ème.	位于巴黎5区的皮埃尔和玛丽-居里大学加希耶校区的扎曼斯基大楼25楼。	3	20
3	DOC16	le 18ème étage de la bibliothèque françois mitterrand, dans le 13ème.	位于巴黎13区的法国国家图书馆密特朗官19楼。	3	20
4	DOC16	le 6ème étage de l'hôtel industriel de dominique perrault, dans le 13ème.	位于巴黎13区的多米尼克-佩罗工业馆7楼。	3	20
5	DOC16	la nuit blanche 2012 permettra aux visiteurs de découvrir la ville lumière d'en haut grâce à 15 belvédères normalement fermés au public.	2012巴黎不眠夜将使参观者可以从15个平时不对外开放的平台发现欣赏巴黎这座光影之城。	3	20
6	DOC16	jk rowling	jk罗琳	3	20
7	DOC16	en effet, jk rowling, qui a créé notre sorcier à lunettes, pourrait se replonger dans l'univers d'harry potter.	事实上,“创造了”我们那位眼镜魔法师的jk罗琳,有可能会写哈利波特的魔法世界里的续集。	3	20
8	DOC16	c'est en 2011 que la saga de nos célèbres sorciers s'est achevée, à l'issue du septième livre intitulé « harry potter et les reliques de la mort » qui a été divisé en deux parties au cinéma.	2011年,第七本《哈利波特和死亡圣器》分为上下两部电影,上映结束之后,我们这著名的哈利波特系列魔法小说完结了。	3	20
9	DOC16	cinq ans après le dernier opus de cette série à succès, jk rowling revient avec une nouvelle œuvre, pour adultes cette fois.	在这成功的系列小说最后一章完结的5年之后,jk罗琳带着她的新作品回归了。这次是给大人看的小说。	3	20
10	DOC16	et la star des librairies a su entretenir le mystère.	这位图书史上的明星会将这光芒延续下去。	3	20
11	DOC14	la joie	欢乐	3	20
12	DOC16	« une bourgade apparemment idyllique mais qui va faire face aux tourments les plus violents ».	“一个表面看起来非常美好的田园小镇,但是它将会面临最猛烈的动荡。”	3	20

Figure 5. Extraction of a "good" TM from a TM produced by "natural" post-edition

No	File name
1	demo2_fr.xml
2	demo2_zh-CN.xml

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <DOC lang="zh-CN">
3 <text>
4 <body>
5 <n =2304> 位于巴黎5区的阿拉伯世界博物馆10楼高级理事会议厅。 </n>
6 <n =2305> 位于巴黎5区的皮埃尔和玛丽-居里大学加希耶校区的扎曼斯基大
7 <n =2306> 位于巴黎13区的法国国家图书馆密特朗官19楼。 </n>
8 <n =2307> 位于巴黎13区的多米尼克-佩罗工业馆7楼。 </n>
9 <n =2308> 2012巴黎不眠夜将使参观者可以从15个平时不对外开放的平台发
10 <n =2310> JK Rowling </n>
11 <n =2311> 事实上,“创造了”我们那位眼镜魔法师的jk罗琳,有可能会写哈
12 <n =2312> 2011年,第七本《哈利波特和死亡圣器》分为上下两部电影,上
13 <n =2313> 在这成功的系列小说最后一章完结的5年之后,JK罗琳带着她的新
14 <n =2315> 这位图书史上的明星会将这光芒延续下去。 </n>
15 <n =1884> 欢乐 </n>
16 <n =2317> “一个表面看起来非常美好的田园小镇,但是它将会面临最猛烈的
17 <n =2318> 在法国,JK罗琳将会在ELLE杂志和TF1电视台介绍她的新书。 </n>
18 <n =2319> 目前,有上百万的读者耐心地等待了这本新书的上市。 </n>
19 <n =2321> 于是之后我问: 为什么会是我? </n>
20 <n =1595> 上帝保佑你们。 </n>

```

Figure 6. Export of a "good" part of a TM

4 Unexpected and costless gains

There are two interesting side effects obtainable without any additional cost: iMAGs can be used to produce high-quality parallel corpora and to set up a permanent task-based evaluation of one or more MT systems

4.1 Production of good quality and “targeted” parallel corpora

Thanks to SECTra_w in-built system of annotation of each translation or post-edition of a segment by a reliability level (from * to *****) and a quality score (0..20), one can extract from the TM associated to a website S a subset verifying any predicate based on levels and scores.

To implement that, we have introduced and implemented into SECTra_w the notion of *selection*.

A selection is defined intentionally (by a predicate) or extensionally (by an explicit list), and can be named, for later recall.

Take for example the TM of the website of Greater Grenoble (La Métro) that contains 2500 web pages, or about 30000 segments. More than half have been pre-translated and post-edited into Chinese for the Shanghai Expo in 2010. We may select a “quite good part” of this TM by creating the selection:

```

TM-lametro-extract-good =
TM_select (lametro, [level=3 &
score >=13 | level=4 & score
>=12 | level=5 & score >=11]).

```

The following example shows an even simpler extraction, from the French-Chinese part of the Demo2 TM associated with iMAG-Doc_Par_jour shown on Figure 5 above. The predicate is

simply [level=3 & score >=13], and its parameters can be directly chosen through the GUI.

The selection obtained can then be exported, as 2 parallel files (source and post-edition) in a simple XML format (Figure 6). SECTra_w also provides additional information (TM, Last updated, Duration of post-editing, post-editor, etc.), and other available download formats (TMX, TXT, and CSV). These data can be used later to “feed” an empirical Moses-based MT system that will become specialized to that website¹⁰.

That possibility is very interesting in the current context. It has been proven that MT systems can be specialized to sublanguages and produce outputs of very high usage value (Chandioux, 1988) (Isabelle, 1987). That means that the outputs are quite readable, and very cheap to post-edit to produce professional quality output.

In recent experiments with a Paris-based multilingual content processing firm, a Moses instance built from a high proportion of a 300K bi-segment TM mixed with a standard parallel corpus extracted from EuroParl (Koehn, 2005) got a BLEU (Papineni et al., 2002) score of about 70%. At this high level, BLEU correlates with usage value: it takes typically 10-15 minutes only to post-edit the equivalent of 1 standard page (250 words, or 400 kanjis), instead of 1 hour to produce a draft translation. But that method works only if a parallel corpus specialized to the sublanguage at hand is available, and that is quite rare in practice.¹¹

The situation is similar if the considered MT system is built by an “expert” method (as TAUM-METEO and then METEO).

For example, there is no available parallel Chinese↔French corpus for e-mails, chats, and short technical notes. Building a parallel corpus from scratch is not an option because of the cost of the operation and the scarcity of translators knowing both languages and the technical terms.

Using an iMAG offers a graceful way to solve that difficulty. Whatever MT systems are available, one can begin without any delay to start the bilingual service needed (a web-based chat, for example), routing messages and documents

through web pages, and using iMAGs to make them accessible (and improvable) in the desired languages. After a while, the TM-S dedicated to the (empirically defined) sublanguage of S will contain enough “good” bi-segments to extract them and use them to build a specialized instance of an MT system (for example, a specialized Moses-S system¹²).

An important point here is that, in order to encourage end users to post-edit, post-editing should be made very simple and user-friendly. One should refrain from transforming it into a debugging environment for some MT systems. That would also go against the principle to be open to as many MT systems as possible.

5 Continuous task-oriented evaluation of one or more MT systems

The second unexpected benefit of online contributive post-editing using SECTra_w as a backend is that it is possible to directly extract from it objective measures, where references are post-edited MT results.

SECTra_w was initially designed to support an MT evaluation campaign organized by France Telecom R&D (Orange Labs). It includes classical scripts to compute BLEU and NIST (Dodgington, 2002), and an original script computing a combination of character-based and word-based edit distances (or semi-distances).

$$\Delta_{\text{comb}}(A, B) = c * \Delta_{\text{char}}(A, B) + (1-c) \Delta_{\text{word}}(A, B) \\ \text{with } 0 \leq c \leq 1.$$

Δ_{char} is computed by the Wagner & Fischer algorithm¹³ (Wagner and Fischer, 1974). To compute Δ_{word} , we consider the (typographic) words of strings A and B as a new set of characters, and apply the same algorithm with a matrix M_{word} such that $M_{\text{word}}[u, v] = \Delta_{\text{char}}[u, v]$. In order to *make the post-editing effort intuitively graspable*, we replace in the W&F Δ_{char} matrix a maximal sequence of N exchanges by N deletions (represented by overstriking and coloring in blue) followed by N insertions (coloring in red).

In evaluation campaigns, one needs to build reference translations, which are produced by expensive professional translators, so that eval-

¹⁰ We are running such an experiment but cannot describe it here for lack of space.

¹¹ Remember: in 2001, Language Weaver (LW) claimed « to be able to produce an MT system overnight » from a large enough parallel corpus. While that was undoubtedly true, LW produced actually only 4 MT systems in 4 years... because parallel corpora corresponding to the translation needs of solvable clients were and are hard to find.

¹² We have built a French-Chinese Moses system for iMAG-LIG, based on 12000 already post-edited segments.

¹³ We use a matrix giving insertion, deletion and exchange costs. Δ_{char} is a distance if all elements are equal to 1, but other values may cause to violate all 3 axioms of distance.

uations are done and redone using the same sets of examples. But, if a website *S* is post-edited in an access language *L*, references are produced continuously as contributors (paid or unpaid, organized or occasional) improve the MT pre-translations or the already available post-editions.

Notice also that there is no need whatsoever to PE all segments. PE is normally done by need. If a segment is badly translated but not important or never read, why improve it? That is one aspect of the “multilingual access” concept that makes it intrinsically cheaper than the traditional translation paradigm.

5.1 Evaluating one or more MT systems

Several MT systems can be called on each source segment in *L1*. When we reconstruct a web page in a target language *L2*, we choose the best (highest score) and most recent post-edition, if any, or one MT output, for example Systran for *En*→*Zh*, or Google Translate for *Zh*→*En*.

We can always compute the available similarity measures (or distances, or semi-distances) between the produced post-edition and each of the MT outputs and each of the other post-editions of the segment. The pseudo-trace presented above illustrates that possibility. In this way, each MT system output can be compared against a reference, which is the result of a post-editing activity that is related to the task at hand. In other words, references are produced naturally and with no additional cost.

If (like we do now) the same MT system *MT-1* is always chosen as initial value of the string to be modified by post-editing, there is a serious risk of a bias in favour of that system, because of the natural tendency of post-editors to modify the pre-translation as little as possible in order to produce a “good enough” translation (post-edition). Then, whatever the measures used, the outputs of *MT-1* will be nearer to the “references” produced by PE than the outputs of the other systems *MT-2*, *MT-3*, ... , *MT-k*.

How to improve on that? A first idea would be to ask *k* humans to post-edit all *k* MT outputs. But that would multiply by *k* the human time taken, and it would clearly be quite unrealistic if one wants to integrate evaluation in a task-related activity without additional cost.

In the future, we plan to choose (automatically) among the *k* possible MT outputs so that each *MT-k* is guaranteed to be used for a fixed propor-

tion of the segments. The simplest way is to “rotate” between systems (choice (*n*) = *n* modulo *k*), so that *n/k* of inputs will be pre-translated by each MT system. It is also possible to “throw the dice”, so that each of *MT-1*, ... , *MT-k* will have 100/*k* % chances to be chosen. There may also be good reasons to give more chances to one MT system, for example to a system being developed and still at the beginning of its “learning curve”. The rotation and controlled random choice methods above can easily be adapted to that idea.

6 Conclusion and perspectives

In this paper we have shown that an interactive Multilingual Access Gateway (iMAG) dedicated to a website *S* (iMAG-*S*) is quite helpful to make *S* accessible in many languages immediately and without editorial responsibility. Visitors of *S* contribute to the continuous and incremental improvement of the most important textual segments, and eventually of all. In this approach, pre-translations are produced by one or more MT systems. To have all (100%) segments post-edited is *not* the goal: it is quite OK if post-edited segments are only those that are important (often accessed) and badly MT-translated.

Continuous use since 2008 on many websites and for several access languages shows that a quality comparable to that of a first draft by junior professional translators is obtained in about 40% of the (human) time, sometimes less, with the condition that contributors post-edit into their native language.

An interesting observation is that post-editors seem to have some kind of personal “expected PE speed” that does not depend on the direction of post-editing. The resulting quality, then, depends only on their expertise in each direction. Note that, although in principle counter-indicated, post-editing from one’s mother tongue may be cost-effective for some situations like Chinese-French in a French firm: acceptable quality at a still reasonable cost can be obtained by PE first the result of Chinese-English MT by a Chinese, and then the result of English-French MT by a French.

We have also shown and illustrated two interesting side effects obtainable without any added cost: an iMAG-*S* can be used to produce a high-quality parallel corpus and to set up a permanent task-based evaluation of one or more MT systems. By nature, the HQ parallel corpus extractable from a TM-*S* is specialized to the sub-

language of the website S. When it becomes large enough after some period of using the iMAG-S (about 10-15000 ‘good’ bi-segments for the sublanguages of classical web sites), it can be used to build an empirical MT system for that sublanguage, and then to improve it incrementally as time goes and new segments are post-edited. Recent experiments in specializing empirical MT systems have shown that remarkably good MT results can be obtained (Rubino et al., 2012). We are running an experiment on French-Chinese that seems to confirm it.

Acknowledgments

This work has been partially funded by UJF, ANR (Traouiero project), L&M (Lingua et Machina) and ANRT.

References

- C-P. Huynh, C. Boitet, and H. Blanchon. 2008. SEC-Tra_w: an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora. *Proc. LREC-08, demonstration session*, 8 p., Marrakech, 27-31/5/08, ELRA/ELDA, ed.
- G. Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. *Proc. HLT 2002*. vol. 1/1: pp. 128-132 (notebook proceedings). San Diego, California. March 24-27, 2002.
- H-P. Nguyen, C. Boitet, and G. Sérasset. 2007. PIVAX, an online contributive lexical data base for heterogeneous MT systems using a lexical pivot. *SNLP- 2007*, 6 p, Bangkok, Thailand, 2007.
- J. Chandioux. 1988. 10 ans de METEO. Traduction Assistée par ordinateur. *Actes du séminaire international sur la TAO et dossiers complémentaires*, OFIL, A. Abbou, ed. Paris.
- K. Papineni, S. Roukos, T. Ward, and W-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, PA.
- M. Daoud, C. Boitet, A. Kitamoto, and M. Mangeot. 2009. Building a Community-Dedicated Preterminological Multilingual Graphs from Implicit and Explicit User Interactions. *Second International Workshop on REsource Discovery (RED 2009)*, co-located with VLDB 2009, Lyon, France, 8 p.
- P. Isabelle. 1987. Machine translation at the TAUM group. Machine Translation: *The State of the Art*, pp. 247–277.
- P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the ACL, demonstration session*, pp. 177–180, Prague, Czech Republic.
- P. Koehn. 2005. Europarl: a Parallel Corpus for Statistical Machine Translation. *Conference Proceedings: the tenth Machine Translation Summit*, pp. 79-86, AAMT, Phuket, Thailand, 2005.
- R-A. Wagner, and M-J. Fischer. 1974. The String-to-String Correction Problem. *JACM 21*: pp. 168-173.
- R. Rubino, S. Huet, F. Lefèvre, and G. Linarès. 2012. Post-édition statistique pour l’adaptation aux domaines de spécialité en traduction automatique, In *Conférence en Traitement Automatique des Langues Naturelles*, pp. 527-534, Grenoble, France.
- W-J. Hutchins, and H-L. Somers. 1992. An introduction to machine translation. *Academic Press*, pp. 175-177, London.

Issues in Incremental Adaptation of Statistical MT from Human Post-edits

Mauro Cettolo †

Christophe Servan ‡

Nicola Bertoldi †

Marcello Federico †

Loïc Barrault ‡

Holger Schwenk ‡

† FBK, Fondazione Bruno Kessler
38123 Povo, Trento, Italy
LastName@fbk.eu

‡ LIUM, University of Le Mans
72085 Le Mans cedex 9, France
FirstName.LastName@lium.univ-lemans.fr

Abstract

This work investigates a crucial aspect for the integration of MT technology into a CAT environment, that is the ability of MT systems to adapt from the user feedback. In particular, we consider the scenario of an MT system tuned for a specific translation project that after each day of work adapts from the post-edited translations created by the user. We apply and compare different state-of-the-art adaptation methods on post-edited translations generated by two professionals during two days of work with a CAT tool embedding MT suggestions. Both translators worked at the same legal document from English into Italian and German, respectively. Although exactly the same amount of translations was available each day for each language, the application of the same adaptation methods resulted in quite different outcomes. This suggests that adaptation strategies should not be applied blindly, but rather taking into account language specific issues, such as data sparsity.

1 Introduction

In this work, we refer to the experimental framework set-up by the MateCat project,¹ which is developing a Web-based CAT tool for professional translators that will integrate new MT capabilities. Among them is what we named self-tuning MT, that is the automatic and incremental adaptation of the MT engine by exploiting user post-edits collected during the life of a translation project.²

¹www.matecat.com

²By *translation project* we mean a set of homogeneous documents assigned to one or more translators.

The main contribution of the paper is to assess the effectiveness of popular SMT adaptation techniques in a real CAT framework, where the supervision is provided through post-edits from professional translators. The methods have been validated in laboratory tests on data collected in a two-day field test, which involved professionals for the translation of English documents to Italian and to German, in the legal domain. This domain represents a relevant sector in the translation industry and is suitable for exploiting SMT, since the information source is sufficiently homogeneous, the language is sufficiently complex, and there is sufficient multilingual data available to train and tune MT models.

The paper is organized as follows. Section 2 lists some of the related works. Section 3 introduces methods used for project adaptation. Section 4 briefly describes the conduct of the field test. Section 5 and Section 6, respectively, introduce the set-up and results of experiments. Section 7 concludes the paper with a discussion on the overall results.

2 Related Work

Our work deals with MT adaptation in general, and incremental adaptation more specifically.

Bertoldi et al. (2012) present an adaptation scenario where foreground translation and reordering models (TM) and language model (LM) of a phrase-based SMT system are incrementally trained on batches of fresh data and then paired to static background models. Similarly, the use of *local* and *global* models for incremental learning was previously proposed through a log-linear combination (Koehn and Schroeder, 2007), a mixture model (linear or log-linear) (Foster and Kuhn, 2007), the filling-up (Bisazza et al., 2011), or via ultraconservative updating (Liu et al., 2012).

Bach et al. (2009) investigate how a speech-to-speech translation system can adapt day-to-day from collected data on day one to improve performance on day two, similarly to us. However, the adaptation of the MT module involves only the LM and is performed on the MT outputs.

On standard machine translation tasks, Niehues and Waibel (2012) compare different approaches to adapt a SMT system towards a target domain using small amounts of parallel in-domain data, namely the backoff, the factored, and the already mentioned log-linear and fill-up techniques; the general outcome is that each of them is effective in improving non-adapted models and none is definitely better than each other, which is the best depending on how well the test data matches the in-domain training data.

This work deals with data selection as well, which is a problem widely investigated by the SMT community, see for example (Yasuda et al., 2008; Matsoukas et al., 2009; Foster et al., 2010; Axelrod et al., 2011). We apply a standard selection technique (Moore and Lewis, 2010), but in a quite different scenario where the task-specific data is extremely small and the generic corpus is actually close to the domain of the task.

3 Adaptation Methods

In this section we describe the techniques employed to adapt our SMT systems, namely data selection and translation, distortion and language model combination.

3.1 Data selection

It has been believed for a long time that just adding more training data always improves performance of a statistical model, e.g. a n -gram LM. However, this is in general true only if the new data is enough relevant to the task at hand, a condition which is rarely satisfied. The typical case is that of a narrow domain, for which a small task-specific text sample can result much more valuable than a very large generic text corpus, coming from sources that may be heterogeneous with respect to size, quality, domain, production period, etc.

The main idea of data selection is to try nevertheless to take advantage of the generic corpus, by picking out a subset of training data that is mostly relevant to the task of interest, which in our case is a specific translation project.

Similarly to (Servan et al., 2012), we first score a generic corpus against a LM trained on a seed of task-specific data, and compute the cross-entropy for each sentence. Then, the same generic corpus is scored against a LM trained on a random sample of itself. The sample size is roughly set equal to the seed corpus. From this point, the difference between task-specific cross-entropy and generic cross-entropy is computed for each sentence. Finally, sentences are sorted on the basis of this score. According the original paper (Gao and Zhang, 2002), this procedure leads to better selection than the simple perplexity sorting.

Now, the best splitting point of the sorted generic corpus has to be determined. The estimation is performed by minimizing the perplexity of a development set on growing percentages of the sorted corpus.

Moore and Lewis (2010) reported that the perplexity decreases when less, but more appropriate data is used (typically reaching a minimum with about 10 to 20% of the generic data). As a side effect, the models become considerably smaller which is an important aspect when deploying SMT systems in real applications.

Note that in our case the selection of parallel text was done by considering only one side of the parallel seed corpus, either the source or the target.

3.2 Adaptation of SMT models

Translation and distortion models: *Fill-up* is a technique for combining translation and distortion models estimated on corpora of different size and content. Initially proposed by Nakov (2008) and then refined by Bisazza et al. (2011), it merges the background phrase table with the foreground phrase table by adding only phrase pairs that do not appear in the foreground table. Only for the translation model, an additional indicator feature signals whether the phrase stems from the foreground or from the background phrase table. We chose the fill-up technique because it performs as good as other popular adaptation techniques (Niehues and Waibel, 2012) but with models that are more compact and easier to tune. It is worth noticing that the fill-up technique investigated by Niehues and Waibel (2012) slightly differs from the one described by Bisazza et al. (2011) in the way the candidate selection is performed.

We also apply a simplified version of the fill-

training	segments (M)	tokens (M)	
		source	target
en→it	1.7	51.1	52.6
en→de	3.2	61.4	67.1

Table 1: Overall statistics on parallel data used for training purposes: number of segments and running words of source and target sides. Symbol M stands for 10^6 .

up, called *backoff*, in which the indicator feature is discarded. Again, the backoff method proposed by Niehues and Waibel (2012) differs slightly in the way the scores of the phrase pairs stemming from the background phrase table are computed.

Language model: As concerns the LM adaptation, we employed the mixture of LMs which consists of the convex combination of one or more background LMs with a foreground LM. The method is available in the IRSTLM toolkit (Federico et al., 2008).

4 Field Test

For each language pair, the field test was organized over two days in which a document had to be translated by four translators. During the first day, for the translation of the first half of the document, translators received suggestions by the baseline MT engines described in Section 6; during the second day, MT suggestions for the second half of the document came from a system adapted to the text of the first day by means of one of the adaptation methods tested in our experiments (Section 6). Translators post-edited machine-generated translations for correcting mistakes and making them stylistically appropriate. The document was selected such that the size of its halves corresponds approximately to the daily productivity of professional translators, that is three to five thousand words.

A report on the field test including an analysis of the productivity of translators has already been published (Federico et al., 2012). Moreover, we performed a preliminary measure of the performance of MT outputs versus the post-edition of each translator. In both cases, pretty large inter-translator differences were observed. Since the limited number of subjects would have led to scores with large variances, we decided to choose

one single representative translator per language pair, postponing analysis statistically more significant to forthcoming field tests involving more translators.

evaluation		segments	tokens	
			source	target
en→it	D0	91	2,960	3,202
	D1	90	3,007	3,421
en→de	D0	86	2,960	2,712
	D1	89	3,007	2,999

Table 2: Overall statistics on test sets used in Day 0 and Day 1 of the field tests.

Ing. pair	name	seed	for test on	%	tokens (M)	
					src	trg
en→it	FGtgt	D0 _{tgt}	D1	10.1	5.1	5.3
	FGsrc	D01 _{src}	D1	9.8	5.1	5.2
	FGtgt	D1 _{tgt}	D0	10.1	5.1	5.3
	FGsrc	D01 _{src}	D0	9.8	5.1	5.2
en→de	FGtgt	D0 _{tgt}	D1	48.1	35.2	32.3
	FGsrc	D01 _{src}	D1	39.6	28.4	26.6
	FGtgt	D1 _{tgt}	D0	38.7	28.3	26.0
	FGsrc	D01 _{src}	D0	21.6	15.3	14.5

Table 3: Statistics of the selected parallel data.

5 Data

Training Data: Training data come from Version 3.0 of the JRC-Acquis collection (Steinberger et al., 2006). Refer to Table 1 for statistics on the actual corpora employed for training.

Evaluation Data: Concerning the evaluation, the document was taken from a motion for a European Parliament resolution published on the EUR-Lex platform in 2012. Statistics on the test documents translated during the field test are reported in Table 2; they refer to tokenized texts. Figures on the source side (English) refer to the texts the users are requested to translate; figures on the target side (Italian/German) refer to the text post-edited by the chosen translator (one for each language pair).³

The data selection described in Section 3.1 was applied to the training corpus. Table 3 provides the amount of data selected for each task. In our

³Although the document to translate is the same for the two language pairs, the segmentation differ due to a language-dependent automatic sentence alignment.

LM	en→it		en→de	
	D0	D1	D0	D1
	PP/OOV	PP/OOV	PP/OOV	PP/OOV
BG	97.5/0.31	93.2/0.27	209.9/1.91	172.9/1.40
FGtgt	73.3/0.54	72.2/0.67	181.4/1.91	147.4/1.43
FGsrc	69.4/0.73	67.6/0.70	166.1/1.95	136.8/1.43
Dn+BG	78.4/0.28	74.3/0.15	201.1/1.84	168.8/1.26
Dn+FGtgt	71.6/0.53	70.9/0.57	170.5/1.84	142.8/1.30
Dn+FGsrc	65.3/0.47	64.1/0.36	156.5/1.88	132.9/1.30
mix(Dn,FGtgt)	76.1/0.53	75.6/0.57	172.3/1.84	145.2/1.30
mix(Dn,FGsrc)	70.7/0.47	69.0/0.36	167.3/1.88	139.2/1.30
mix(Dn+FGtgt, BG)	80.8/0.28	78.1/0.15	185.5/1.84	167.8/1.26
mix(Dn+FGsrc, BG)	80.3/0.28	77.0/0.15	186.8/1.84	167.6/1.26
mix(Dn, FGtgt, BG)	66.8/0.28	64.7/0.15	169.3/1.84	146.3/1.26
mix(Dn, FGsrc, BG)	65.3/0.28	62.5/0.15	165.4/1.84	144.1/1.26

Table 4: Perplexity (PP) and out-of-vocabulary rate (OOV) of D0 and D1 on different 5gr LMs.

experiments, D0 and D1 alternatively played the role of development and test set. The seed for the selection was either the target side of the development set (Dn_{tgt} , $n=0,1$) or the concatenation of the source side of both the development and test set ($D01_{src}$); we name $FGtgt$ and $FGsrc$ the selected corpus and the models trained on it in the two cases.

The table also provides the percentage of data selected, computed with respect to the target side. The optimal splitting was performed by minimizing the perplexity of the target side of the development set.

6 Experiments

Lab test experiments have been performed on data sets described in Section 5. Performance are provided in terms of BLEU and TER, computed by means of the `MultEval` script implemented by Clark et al. (2011), and of GTM.⁴ For statistical significance, p-values were calculated via approximate randomization for adapted systems with respect to the baselines and are reported in Tables 5 and 6 whenever not larger than 0.10.

The SMT systems have been built upon the open-source MT toolkit Moses (Koehn et al., 2007). The translation and the lexicalized re-ordering models are trained on the available parallel training data (Table 1); 5-gram LMs smoothed through the improved Kneser-Ney tech-

nique (Chen and Goodman, 1999) are estimated on the target side via the IRSTLM toolkit (Federico et al., 2008). The weights of the log-linear interpolation model have been optimized by means of the Margin Infused Relaxed Algorithm (MIRA) process (Hasler et al., 2011) provided within the Moses toolkit.

Various models have been built by means of the methods described in Section 3. Here the list of acronyms and corresponding meaning used in the rest of the paper. Note that whenever “data selected” is mentioned, we refer to the application of the procedure described in Section 3.1 with the training data playing the role of *generic corpus* and the portion of the document translated during either the first day (D0) or the second day (D1) that of *seed corpus*:

BG: background model, trained on the whole training data

Dn+BG: model trained on the concatenation of Dn (either D0 or D1) and training data

FGtgt: model trained on data selected using the target side of either D0 or D1 as seed corpus

Dn+FGtgt: model trained on the concatenation of the target side of either D0 or D1 and FGtgt

FGsrc/Dn+FGsrc: similar to FGtgt/Dn+FGtgt, but the selection is made using the concatenation of the source side of both D0 and D1 as seed corpus

⁴<http://nlp.cs.nyu.edu/GTM>

TM=BG LM	en→it						en→de					
	D0			D1			D0			D1		
	BLEU	TER	GTM	BLEU	TER	GTM	BLEU	TER	GTM	BLEU	TER	GTM
BG	47.9	34.9	73.6	46.8	34.7	74.3	32.2	58.4	60.1	37.4	49.1	65.5
Dn+BG	50.2 [▲]	33.2 [▲]	75.0	48.1 [△]	34.0 [△]	74.9	32.4 [◇]	57.1 [△]	59.7	37.9 [◇]	48.6 [△]	65.7
Dn+FGtgt	46.4	36.1	72.5	47.8	34.5	74.4	32.6 [◇]	56.5 [△]	59.7	37.9 [◇]	48.1 [△]	65.5
Dn+FGsrc	47.6	34.8	73.6	48.2	35.0	74.2	33.0 [◇]	56.2 [△]	60.6	37.9 [◇]	47.9 [△]	65.5
mix(Dn,FGtgt)	45.7 [◇]	35.2	73.5	46.8	34.7	74.3	33.0 [◇]	56.0 [△]	60.3	36.1	49.0 [◇]	65.4
mix(Dn,FGsrc)	47.0	34.4	74.4	47.3	34.5	74.2	33.7 [△]	55.4 [▲]	60.8	36.2	48.8 [△]	65.4
mix(Dn+FGtgt,BG)	49.8 [▲]	33.0 [▲]	75.3	47.7 [△]	33.8 [△]	75.0	33.4 [△]	55.3 [▲]	59.8	36.0	49.1	64.3
mix(Dn+FGsrc,BG)	48.8	33.6 [△]	74.8	47.4	34.1 [◇]	74.7	33.8 [△]	54.6 [▲]	60.3	35.8	49.5	64.1
mix(Dn,FGtgt,BG)	47.6	34.0	74.3	48.5 [◇]	33.7	74.9	33.3 [◇]	56.4 [△]	60.5	36.8	48.1 [△]	66.2
mix(Dn,FGsrc,BG)	48.5	33.6	75.4	48.6 [◇]	33.2 [△]	75.1	32.8 [◇]	56.9 [△]	60.5	36.7	48.2 [△]	66.0

Table 5: Performance on D0/D1 of systems with LMs adapted on D1/D0. Symbols [▲], [△] and [◇] near to BLEU and TER scores indicate that adapted models outperform BG with p-values not larger than 0.01, 0.05 and 0.10, respectively.

`mix()`: mixture of LMs (linear interpolation)

`fillup()`: fill-up of TMs

`backoff()`: backoff of TMs

It is worth noticing that using the source side of the test set for data selection (systems `FGsrc` and `Dn+FGsrc`) can be ambivalent. On the one hand the system can be penalized since its LM is estimated on the target side of the selected parallel data; on the other hand it can be rewarded since the seed corpus includes the actual text to translate, and hence the selected data could be more appropriate.

First of all, the quality of LMs was assessed in terms of perplexity (PP) and out-of-vocabulary (OOV) rate. Indeed, in the computation of PP of a text with respect to a given LM, the presence of OOV words is accounted by adding a fixed penalty for each OOV occurrence; nevertheless, we think useful to provide even explicit OOV values for the sake of completeness. Scores are provided in Table 4: they refer to D0 and to D1 and are computed on the baseline LM (BG) or on LMs adapted in various ways to the other portion of the field test document (D1 or D0).

In general, adapted models always improve the PP of the baseline LM, while the OOV decreases provided that the whole training text is also used to train the model. More specifically:

- data selection is effective: with reference to `FGtgt` and `FGsrc` rows, whatever the seed, the

PP of Dn on the selected data always improves over the baseline, from 15% up to 30% depending on the target language and on the test set; of course, the OOV rate worsens because the lower amount of training data

- the selection on the concatenation of the source side of the development and evaluation sets is more effective than the selection made only on the target side of the development set: compare paired rows including `FGsrc` and `FGtgt`

- the linear interpolation of LMs gives contrasting results: from one side, `mix(Dn, FGsrc, BG)` allows the lowest PP on Italian and very competitive on German; from the other, when it is applied to Dn and `FGtgt/src` it fails with respect to the naive concatenation

- the use of the development set in LM training yields a significant improvement of the OOV rate, especially for D1 (from 0.27% to 0.15% for Italian, from 1.40% to 1.26% for German); at the same time, `Dn+BG` row shows a significant PP improvement over the baseline only for Italian: this means that D0 and D1 are alike for that language pair, less for English-German where consequently the adaptation could be more problematic.

6.1 MT results

MT experiments have been conducted either by varying the LM and keeping fixed the baseline TM (Table 5) and by consistently pairing the adapted models (Table 6). Baseline MT system uses BG

adapted LM/TM	en→it						en→de					
	D0			D1			D0			D1		
	BLEU	TER	GTM	BLEU	TER	GTM	BLEU	TER	GTM	BLEU	TER	GTM
BG	47.9	34.9	73.6	46.8	34.7	74.3	32.2	58.4	60.1	37.4	49.1	65.5
Dn+BG	49.9	32.9 [◇]	76.0	49.8 [▲]	33.3 [◇]	75.8	32.6 [◇]	57.4 [△]	60.0	37.6 [◇]	48.4 [△]	65.6
Dn+FGtgt	45.5	35.8	72.9	49.3 [△]	33.2	75.5	31.9	58.0 [◇]	59.6	38.1 [△]	48.2 [△]	65.9
Dn+FGsrc	48.3	33.7	74.3	49.6 [△]	33.2	75.5	31.9	58.90	58.3	37.9 [△]	48.4 [△]	66.1
mix/fillup(Dn,FGtgt)	45.2	36.7	73.1	46.9	34.8	75.0	30.6	58.5	58.5	35.1	50.80	64.2
mix/fillup(Dn,FGsrc)	45.8	35.6	73.7	48.4	33.9	75.4	31.3	59.0	58.6	35.2	51.30	64.3
mix/fillup(Dn+FGtgt,BG)	50.3 [△]	32.5 [▲]	76.0	49.4 [▲]	33.3 [△]	75.6	31.6	58.4	59.4	38.5 [▲]	47.8 [▲]	65.7
mix/fillup(Dn+FGsrc,BG)	48.8	33.4 [◇]	75.0	49.5 [▲]	32.3 [▲]	76.1	31.1	59.0	58.9	37.9 [△]	48.9 [△]	65.7
mix/backoff(Dn+FGtgt,BG)	50.5 [▲]	32.4 [▲]	76.1	49.1 [△]	33.4 [◇]	75.4	33.0 [△]	56.7 [▲]	60.4	38.9 [▲]	48.1 [▲]	66.0
mix/backoff(Dn+FGsrc,BG)	49.2	32.9 [△]	75.1	49.1 [△]	32.3 [▲]	76.0	31.8	59.1	59.6	38.3 [△]	48.4 [△]	65.9
mix/fillup(Dn,FGtgt,BG)	46.6	35.5	73.7	49.0 [◇]	33.3	76.0	31.6	58.3	58.9	36.1	51.0	63.6
mix/fillup(Dn,FGsrc,BG)	48.1	34.4	75.2	49.4 [△]	32.9 [◇]	75.8	30.3	59.7	58.8	36.5	50.6	64.3

Table 6: Performance on D0/D1 of systems with models adapted on D1/D0. Symbols [▲], [△] and [◇] near to BLEU and TER scores indicate that adapted models outperform BG with p-values not larger than 0.01, 0.05 and 0.10, respectively.

models; its results are replicated in the first row of the two tables for the sake of readability.

The first set of experiments aimed at isolating the contribution of adapted LMs, a fortiori since adaptation often involves just the LM. The symmetric experiments where only the TM is changed are less informative since the lack of LM support prevents improvements by the TM to emerge; therefore, they are not presented.

Adapted LMs: in many cases, Italian adapted LMs allows to outperform the performance of the baseline system, whereas no method yield significant improvements on both days in the English-German task; this should be due to the degree of similarity between D0 and D1: pretty high for English-Italian, quite low for English-German, as stated before in comments to Table 4.

For the English-Italian pair, in general the better the PP and OOV values are, the better the translation is. In fact, the low PP of LMs built over Dn and FGtgt/src only, does not yield good MT scores, because the high OOV rates. The naive concatenation of Dn and BG provides surprisingly good performance, matched only by the mix(Dn+FGtgt/src,BG) LMs; the interpolation of the three LMs, which gave the best PP, keeps its promise only on D1. Differently than PP, data selection on the source side of D0 and D1 not always overcomes that on the target side of the de-

velopment set.

Concerning the English-German pair, the naive concatenation of Dn and BG does not improve nor hurt baseline performance. Again mix(Dn+FGtgt/src,BG) outperforms the baseline, but limited to D0. On D1 the only effective method is the concatenation of development and selected data (Dn+FGtgt/src).

A common outcome regards the unreliability of the Dn model: whenever it is combined with other LMs as it is (mix(Dn, ...)), effects are mostly negative compared to concatenation; this is due to the small amount of data used for its training (about 3,000 words).

Adapted Models: the different effectiveness of methods on the two language pairs observed by changing only the LM is confirmed when adapted LMs and TMs are consistently paired: most techniques are effective on English-Italian with the added value guaranteed by the reciprocal support of models, while controversial results characterize the English-German pair.

For the favorite pair, adapted systems significantly outperform the baseline provided that the whole training data is somehow used in building models. It deserves mentioning the excellent performance of the models built over the naive concatenation of the development and training data (Dn+BG). The best systems seem to

be those combining $(Dn+FGtgt, BG)$ for D0, $(Dn+FGsrc, BG)$ for D1: the fact that the references of D0 and D1 are post-edits and are used both for evaluation and for building adapted models could explain that apparently incoherent behavior. Again, the use of the model built on just Dn negatively affects performance.

On English-German task, the only good-performing technique on both days is the $mix/backoff(Dn+FGtgt, BG)$, while the naive concatenation $Dn+BG$ slightly improves some scores and does not affect the others. Evidently, D0 and D1 are too different to allow models adapted on one of them to well represent also the other.

An outcome shared by the two tasks is that $backoff$ is a bit more effective than $fillup$, probably due to the difficulty in properly setting the weight of the additional indicator feature of the latter method.

7 Discussion

The experiments with Italian and German translations, although performed on the same source texts and by applying the same adaptation methods, result in quite different outcomes.

We try now to summarize the main issues and to sketch possible explanations and directions we will investigate in the future to overcome them.

Data selection. The same amount of seed data (3,000 words) does not work equally for German and Italian. While for Italian, around 10% of the training data were selected by seeding with D0 and D1 texts, between 20%-48% were selected for German. Data selection relies on similarity scores computed using small language models estimated disregarding infrequent words (Moore and Lewis, 2010). Given its highly inflected language, it is likely that for German the large majority of project specific words in the seed are singletons, so that the corresponding LM loses most of its specificity. Alternative ways to explore, specifically for highly inflected languages such as German, could be to remove word inflection during data selection, e.g. by stemming words, and to work with low-order n -grams, e.g. 1-grams and 2-grams.

LM adaptation. All the tested LM adaptation methods provided improvements in terms of perplexity and OOV rate (Table 4), both on the Italian

and German translations. Such improvements reflected for some adaptation methods in better MT scores (Table 5) for the English-Italian direction, but not always for the English-German direction. It is worth noticing that the simple concatenation of training and adaptation data performs better than more refined and probably too aggressive adaptation approaches. The inconsistent behavior of perplexity and translation scores for both translation directions can be explained by the fact that adapted LMs basically boost the probability of subsets of target words, that should likely occur in the reference test translations, thus giving better perplexity values and OOV rates. However, if the same target words are not reachable through the translation model, the advantage provided by the adapted LM vanishes. This mismatch becomes even more relevant when adapted translation models are employed, too.

TM adaptation. The use of only adapted LMs showed significant improvements to slight degradations in performance, according to the considered method, translation directions, and adaptation and testing sets. The addition of adapted translation models (Table 6) further widened the range of outcomes and, mostly important, does not show to be additive with language model adaptation. In fact, language model adaptation configurations that perform best do not seem to combine well with some translation model adaptation methods, especially for English-German. In fact, the most consistent behavior across all languages and data sets is shown by a specific configuration ($mix/backoff(Dn+FGtgt, BG)$), whereas very similar set-ups show mixed behaviors. A possible reason for this may be the overfitting of the TM on the adaptation data. In particular, as for each English word more German surface forms may correspond than for Italian, biasing the TM towards the observations of the adaptation data can hurt the overall quality of adapted models. Concerning data selection, the problem with German seems that the seed is not large enough to properly characterize the narrow domain of the document. Hence, in such a case, only soft adaptation methods appear adequate and safe.

As future work, we plan to investigate both on data selection in case of small seeds and on less

aggressive adaptation methods for inflected languages, such as biasing the translation model only at the lexical rather than phrase level and to generalize over different word inflections. Moreover, other field tests will be carried out in order to collect further post-edited translations.

Acknowledgments

This work was supported by the MateCAT project, which is funded by the EC under the 7th Framework Programme.

References

- Axelrod, A., X. He, and J. Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *EMNLP*, Edinburgh, UK.
- Bach, N., R. Hsiao, M. Eck, P. Charoenpornasawat, S. Vogel, T. Schultz, I. Lane, A. Waibel, and A. W. Black. 2009. Incremental adaptation of speech-to-speech translation. In *NAACL HLT: Short Papers*, Boulder, US-CO.
- Bertoldi, N., M. Cettolo, M. Federico, and C. Buck. 2012. Evaluating the learning curve of domain adaptive statistical machine translation systems. In *WMT*, Montréal, Canada.
- Bisazza, A., N. Ruiz, and M. Federico. 2011. Fill-up versus interpolation methods for phrase-based SMT adaptation. In *IWSLT*, San Francisco, US-CA.
- Chen, S. F. and J. Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 4(13):359–393.
- Clark, J., C. Dyer, A. Lavie, and N. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *ACL*, Portland, US-OR.
- Federico, M., N. Bertoldi, and M. Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Interspeech*, Melbourne, Australia.
- Federico, M., A. Cattelan, and M. Trombetti. 2012. Measuring user productivity in machine translation enhanced computer assisted translation. In *AMTA*, Bellevue, US-WA.
- Foster, G. and R. Kuhn. 2007. Mixture-model adaptation for SMT. In *WMT*, Prague, Czech Republic.
- Foster, G., C. Goutte, and R. Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *EMNLP*, Cambridge, US-MA.
- Gao, J. and M. Zhang. 2002. Improving language model size reduction using better pruning criteria. In *ACL*, Philadelphia, US-PA.
- Hasler, E., B. Haddow, and P. Koehn. 2011. Margin infused relaxed algorithm for Moses. *The Prague Bulletin of Mathematical Linguistics*, 96:69–78.
- Koehn, P. and J. Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *WMT*, Prague, Czech Republic.
- Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *ACL Companion Volume Proc. of the Demo and Poster Sessions*, Prague, Czech Republic.
- Liu, L., H. Cao, T. Watanabe, T. Zhao, M. Yu, and C. Zhu. 2012. Locally training the log-linear model for SMT. In *EMNLP*, Jeju, Korea.
- Matsoukas, S., A.-V. I. Rosti, and B. Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *EMNLP*, Singapore.
- Moore, R. C. and W. Lewis. 2010. Intelligent selection of language model training data. In *ACL Short Papers*, Uppsala, Sweden.
- Nakov, P. 2008. Improving English-Spanish statistical machine translation: experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *WMT*, Columbus, US-OH.
- Niehues, J. and A. Waibel. 2012. Detailed analysis of different strategies for phrase table adaptation in SMT. In *AMTA*, San Diego, US-CA.
- Servan, C., P. Lambert, A. Rousseau, H. Schwenk, and L. Barrault. 2012. LIUM’s statistical machine translation systems for WMT 2012. In *WMT*, Montréal, Canada.
- Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufi, and D. Varga. 2006. The JRC-Acquis: a multilingual aligned parallel corpus with 20+ languages. In *LREC*, Genoa, Italy.
- Yasuda, K., R. Zhang, H. Yamamoto, and E. Sumita. 2008. Method of selecting training data to build a compact and efficient translation model. In *IJCNLP*, Hyderabad, India.

The ACCEPT Post-Editing Environment: a Flexible and Customisable Online Tool to Perform and Analyse Machine Translation Post-Editing

Johann Roturier, Linda Mitchell, David Silva

Symantec Ltd.

Ballycoolin Business Park

Blanchardstown, Dublin 15

Ireland

{johann.roturier,linda.mitchell,david.silva}@symantec.com

Abstract

This paper presents an online environment aimed at community post-editing, which can record and store any post-editing action performed on machine translated content. This environment can then be used to generate reports using the standard XLIFF format with a view to provide stakeholders such as machine translation providers, content developers and online community managers with detailed information on post-editing actions. This paper presents in detail the functionality available within the environment as well as the design choices that were made when creating this environment. Preliminary usability feedback received to date suggests that the feature set is sufficient to perform community post-editing.

1 Introduction

Machine translation (MT) systems are increasingly used to produce rough translated versions of documents that may be reviewed and possibly modified by post-editors in order to produce improved versions. For instance, Dell and Welocalize recently announced an MT-based localisation program for the translation of Web content using 27 MT engines.¹ Thanks to this deployment, machine-translated documents (such as technical support articles) are made available to end-users who may glean useful information if their knowledge of the source language is not sufficient and if the quality

of the MT output is sufficient. In some cases, however, previous studies have shown that the quality of MT output may not be sufficient for the output to be found comprehensible by end-users (Roturier and Bensadoun, 2011), especially when the source text is uncontrolled (Mitchell and Roturier, 2012). This is especially true when source content is user-generated, which is why the ACCEPT project aims at developing new technologies designed specifically to help MT work better in an online community environment.²

Publishing documents that are difficult or impossible to understand defeats the purpose of publishing documents in the first place, so post-editing these documents before (or just after) they have been published is often considered as an important step in document localisation production workflows (Flournoy and Duran, 2009). Providing post-editors with the right environment to perform this activity has received a lot of attention lately since post-editing is a very different task from translation. This paper, which presents an online environment aimed at community post-editing, is divided as follows: Section 2 reviews existing post-editing environments by highlighting missing functionality from a community perspective. Section 3 presents the choices that were made when designing the ACCEPT environment and Section 4 describes its various functionality. Section 5 presents the results of a small evaluation study conducted with the help of an online survey, seeking to elicit feedback from users of the environment. Finally conclusions and suggestions for future work are presented in Section 6.

¹<http://www.welocalize.com/dell-welocalize-the-biggest-machine-translation-program-ever/>

²<http://www.accept-project.eu/>

2 Related work

MT technology is increasingly used by Language Service Providers (LSP), as revealed by a 2009 TAUS market survey, which showed that 40% of the surveyed LSPs already used MT, with a majority of the remaining 60% indicating that they were considering an MT integration in their processes in the next two years.³ The reasons for the increase in MT technology adoption are varied. One obvious reason concerns the productivity gains in the translation industry reported by several studies, such as Plitt and Masselot (2010), de Almeida and O'Brien (2010), Guerberof Arenas (2012) and (Green et al., 2013). Another reason is related to the increasing ubiquity of machine translation tools (such as Google Translate) and the proliferation of tools providing an environment where machine-translated text can be improved. Such environments include desktop-based standalone tools, web-based generic environments and web-based dedicated environments.

Desktop-based tools include generic computer-aided tools (CAT) or translation environment tools (TenT) that have been enhanced to support the post-editing of machine-translated content. Such tools tend to be aimed at professional translators, rather than community users, so they will not be reviewed any further. Dedicated standalone post-editing tools aim at studying the work of post-editors, for example by recording post-editing actions thanks to keylogging or eye-tracking software. Examples of such tools include Translog II (Carl, 2012), PET (Aziz et al., 2012) or iOmegaT⁴. Our work differs from these standalone tools since the reports generated by our environment are more concise and provide a summarising overview of keylogging actions per revision, per segment, per task for each participant.

Web-based dedicated tools are tightly integrated with the platform where the source and target content is created and published. An example of such an environment is the wikiBABEL platform (Kumaran et al., 2008), which provides a user interface and linguistic tools for collaborative correction of the rough content by a community of users, thus helping the creation of improved content in the

target language. While the ACCEPT environment also targets user communities, it is not dedicated to a single community of users (such as Wikipedia). As described in section 3, our environment is currently available by logging to a portal, but its architecture is sufficiently flexible that it could in theory be used in any content management environment.

Web-based generic tools allow users to log on to an online system to post-edit machine-translated content. Examples of such environments include MateCat⁵ and TransCenter (Denkowski and Lavie, 2012). These tools allow translators to log on to a web-based translation editor to view sentences in a simple, easy-to-follow grid format. Both tools allow project data to be exported in HTML or comma separated value (CSV) format. Our work differs from these environments with respect to two main points: the ACCEPT environment is aimed at community users so a grid format to display source and target texts does not seem appropriate. As described in Section 4, our user interface focuses on the target text. Our environment also supports the export of data in XLIFF format instead of CSV format to maximise interoperability. Another tool that falls in this area is CASMACAT, which is a sophisticated tool that aims at investigating the integration of technology in translation using logging and eye-tracking technology (Elming and Bonk, 2012). The ACCEPT tool differs from CASMACAT in many aspects, including the technology used for the client application. While CASMACAT uses both HTML5 and JavaScript in its client application, the ACCEPT client application is written in JavaScript and uses JQuery libraries.

Finally it is worth mentioning the Microsoft Collaborative Translation Framework,⁶ which is a hybrid web-based method enabling specific users to submit corrections and to retrieve translation candidates.⁷ While this framework provides users with the ability to retrieve contributed segments in real time, it does not allow for any other informa-

³<https://www.taus.net/reports/lsp-in-the-mt-loop-current-practices-future-requirements>

⁴<http://try-and-see-mt.org/>

⁵http://www.matecat.com/wp-content/uploads/2013/01/MateCat-D4.1-V1.1_final.pdf

⁶<http://blogs.msdn.com/b/translation/archive/2010/03/15/collaborative-translations-announcing-the-next-version-of-microsoft-translator-technology-v2-apis-and-widget.aspx>

⁷<http://msdn.microsoft.com/en-us/library/hh847650.aspx>

tion related to the post-edited actions (such as post-editing time) to be retrieved.

3 Design choices

Unlike other tools, the ACCEPT environment is divided into four distinct parts: a database, a server-side Application Programming Interface (API), a Web-based application (portal) and a client application where the actual post-editing work can be performed. While the client application is also Web-based, it differs from the Web-based portal because it is written in JavaScript (using JQuery libraries). This means that the client application can be integrated into any third-party Web application without requiring users to log on to the ACCEPT portal. This approach presents the advantage of bringing users closer to the actual content that should be post-edited.

The server-side API is used to create projects and tasks and upload and download data thanks to forms that are made available to specific users of the Web-based portal, known as project administrators (i.e. users that have sufficient rights to create projects). The API is also used to save any post-editing action that takes place in the client application. It is difficult to identify where machine translation fails from a post-editing perspective. By storing the post-editing actions (corrections of machine translated output and UI interactions) taken by multiple users in a central online location (the database), weaknesses of the MT systems can be quickly identified, especially since this information can be made available to project administrators in real-time. A crucial characteristic of the the ACCEPT environment is related to the way user activity is recorded (e.g. time spent, keys pressed): the recording is purely limited to what happens within the client application. This means that any activity performed outside of the ACCEPT environment (e.g. looking up a word definition or frequency using a search engine) will be ignored to respect the user's privacy. Further architecture and data management details are provided in the next sections.

3.1 Managing projects

The ACCEPT environment allows project administrators to create and manage post-editing projects. It is expected that project administrators have some community management expertise or

responsibilities in order to create projects in which their community members may be interested in participating. It is the responsibility of the project administrators to decide who to invite to work on a project. The content that project administrators may upload for post-editing may range from short community contributions (such as forum or blog posts) to longer organisation-related documents such as training material (say, produced by a non-governmental organisation). Projects can be created to collect user edits and to possibly study those edits in detail thanks to the additional information that gets recorded during a post-editing session (such as time spent, types of keys pressed, etc.). This information can then be used by project administrators to identify participants who have contributed the most translations and/or the best translation quality to a project with a view to rewarding them using existing community rewarding schemes. In order to cater for multiple post-editing scenarios, the actual functionality of the client application is defined during project creation. To provide project administrators with as much flexibility as possible, the following interface configuration options are available during project creation:

- How is the post-editing task going to be conducted? In a monolingual or bilingual manner? If it is conducted in a bilingual manner, source segments may be viewed by users. Otherwise, source segments may be hidden. Default post-editing guidelines are also affected by this choice. A specific setting allows project administrators to decide whether users can override this project-level configuration (e.g. while a project may be configured in a monolingual manner, users can still view the source if they want to by clicking on a switch).
- Should translation options be used during a project? When this option is selected, alternative translation options provided by the MT system may be displayed to users.
- Should a specific feedback question be active? If so, which question/values should appear? When this field is used, the first option appears as the title of a drop-down list, followed by possible values. This allows project administrators to elicit feedback from users

on various aspects of the tasks (e.g. quality of the source, quality of the MT output, etc.)

- Should a post-task question be active? If so, which question should appear? Should a post-project link to a survey be present?
- What is the language pair of the project? If the language pair is English > French, the User Interface will be displayed in French. Currently, the following User Interface and project languages (i.e. languages in which content can be post-edited) are supported, but more could be easily added in the future: English, French and German. This selection also influences the language resources used by the spelling and grammar checker.
- Which user(s) should be invited to take part in this project and what text should be used to invite users to the project?

Once invited users have registered with the ACCEPT portal, they are presented with a list of tasks to work on. Once these tasks are completed, they disappear from their project page. Users can start working on a task by clicking its task ID. Once the task ID is clicked, the post-editing window appears based on the configuration that was specified by the project administrator. All user actions are saved automatically (e.g. segment-level comments, segment changes), but users are able to use the Undo and Redo functionality when editing segments. Users can close the window even if a task is not completed. This stops the global time count until the window is opened again. A task can be closed for good at any moment by the user (after being prompted to confirm their choice).

Once a project is created, data that should be post-edited by human reviewers can be imported into the system, as described in the next section.

3.2 Uploading data

The data format that is currently supported is based on a simple JSON format,⁸ which must contain the following data:

- *text_id*: a string corresponding to a unique string identifier (e.g. hash value of the source text). A file with a *text_id* that is already in use in a given project can not be uploaded.

⁸<http://www.json.org/>

- *src_sentences*: an array of objects, which are pairs of sentences, in the form: “*text*”: “*sentence*” (where *sentence* is a string)
- *tgt_sentences*: an array of objects, each containing a sentence pair, in the form: “*text*”: “*sentence*”. Each object may also contain an optional options pair, in the form: “*options*”: *option_array* (where *option_array* is an array of objects). Each object in the *option_array* should contain a pair of tokens (in the form “*context*”: “*token*”, where *token* corresponds to a substring from a target sentence) and a *values* pair which should contain an array of objects. Each of these objects should contain an alternative phrase pair, in the form “*phrase*”: “*token*”.

The number of objects in *src_sentences* must be equal to the number of objects in *tgt_sentences*. The final JSON format may also contain metadata, such as a contact email address, an MT tool name and tool ID, a source language code and a target language code.

The final JSON format may also contain an optional *tgt_templates* array to influence the display of the target sentences in the post-editing window. Templates were developed to define the layout of the post-editing tasks and how they are displayed in the editor. Standard DIV elements may be included within the *tgt_templates* array and each DIV element may contain any CSS style information within or around it. Each DIV corresponds to one segment in the editor, which means that the display of a paragraph can be defined individually. Each DIV element must include a *@TARGET@* sequence, which defines the position of the respective target sentence in the target text. The number of segments in *tgt_templates* needs to match the number of segments in *src_sentences* and in *tgt_sentences* in order for the templates to be mapped correctly to the segments to be edited. To build a paragraph, for example, the sequence *style=“display:inline”* can be included in front of the placeholder for the actual segment. Special characters must be encoded (e.g. *<* as *%3Cdiv*) since the server only accepts characters inside the ASCII character set.⁹ An example of such a DIV is shown below:

⁹http://www.w3schools.com/tags/ref_urlencode.asp

```

    "tgt_templates": [ { "markup": "%3Cdiv
style=\\"display%3Ainline;\\"%3E@TARGET@
%3C/div%3E%3E&nbsp;" },
    { "markup": ... } ],

```

When no template is specified, the default behaviour is adopted and displayed, which means that the post-editing task on the left of the editor contains no paragraph. The default style for a DIV is `style="display:block"`.

3.3 Designing the client application

While the actual functionality available in the client application is described in section 4, the present section focuses on technical design choices. When designing the client application of the ACCEPT post-editing environment, the focus was on developing a portable application purely written in JavaScript instead of relying on a standard Web application. A standalone Web application would have required users of an existing community to switch environments to perform post-editing tasks. The jQuery library was selected to build this portable application as a plug-in.¹⁰ This library was selected because it is fast and widely used. It also has useful utility functions, good documentation and a large community. It offers cross-browser compatibility and full support for CSS3 selector specification. When it is used in conjunction with the jQuery UI library, CSS styles can be inherited and easily changed.¹¹ The logic around the technical aspect of the plug-in is based on collecting information from an HTML DOM object that self-configures the plug-in.

3.4 Recording post-editing actions

In order to assess whether MT or post-editing was effective, analysing the actions performed or the time spent by a given user on a given task may be extremely informative. The ACCEPT environment records actions and time in the following manner.

1. When a user clicks on any task link, the post-editing window opens.
2. When the post-editing window opens, the global time count for the task starts. This action is captured in a phase called *start_pe*.
3. When a user clicks on any segment in the target text (say, segment 1), the target segment appears in the editing window.
4. When the user takes any action in the editing window, a time-stamp is recorded for the first revision of the current segment. This action is recorded in a phase called *r1.1* (where *r* stands for revision, *1* for segment 1, and *1* for revision 1).
5. The user now leaves the current segment, by clicking on another segment (say segment 3), and starts editing it. There is a new time-stamp for this new segment, at revision 1. Keystrokes are recorded again.
6. The user is not satisfied with segment 1, and clicks on segment 1. There is a new time-stamp for segment 1 at revision 2. Editing actions are recorded.
7. The user clicks on segment 2. No further actions are taken by the user. No extra information is recorded.
8. The user is satisfied with the result of this task, clicks the *Complete Task* button, and is asked to confirm. If the confirmation is positive, the status for the document is *FINISHED*. In this case, the task disappears from the overview page and the global time count for the task stops. Otherwise, the status of the task is *UNFINISHED*. In that case, the task will still be displayed on the project page. The global time count for the task stops until the task is re-opened.

During the post-editing process, users often behave in unforeseen ways or come across unforeseen issues. In most existing solutions, user actions and interactions can only be investigated once the post-editing process has been completed. The ACCEPT environment provides instant access to user progress, so that project administrators can identify potential problems and solve them on the spot without losing valuable time and data. This real-time recording of the users' progress facilitates a more efficient project management which can also be used for rewarding users according to their progress automatically. The next section focuses on the format of the report used to export user activity.

¹⁰<http://jquery.com/>

¹¹<http://jqueryui.com/>

3.5 Generating activity reports

Actions and interactions from multiple users are collected and grouped together in real-time. A standard format (XLIFF) is used for the export of information, thus maximising interoperability.¹² Figure 1 presents the mapping used to group a data point (such as the number of arrow keys pressed in a revision of a given segment) in an XLIFF *count* element with a *count-type* attribute whose value has a custom value, *x-arrow-keys*.

As shown in Table 1, multiple XLIFF elements are used to represent post-editing activity data. The actual target text entered by a user when modifying a machine-translated segment is present in the *body* element in *target* elements of *trans-unit* elements. Each revision is saved as a distinct phase labeled with a unique *phase-name* attribute. This attribute value can then be cross-referenced with a *count* element and a *phase* element that have the same attribute value. These *count* and *phase* elements contain metadata information about the activity performed by the user. For instance, the number of arrow keys pressed or any comment the user may have made on the quality of the original machine-translated segment can be represented using these elements.

Project administrators can then export post-editing activity data at the user-, document- or project -level by generating XLIFF files, such as the one shown in Figure 1:

Figure 1 shows the activity one user performed on one task. A task is mapped to an XLIFF *file* element, which contains a *header* element and a *body* element. The *body* element is used to capture all of the texts that were used and produced during the post-editing process, including the source text, the original machine-translated text and any revision that may have been produced. The *header* element contains detailed information on the actual actions that occurred during the post-editing process (e.g. types of keys being pressed, number of times the text checker was used, alternative options used, etc.).

4 Client Functionality

The guiding principle followed when selecting functionality for the plug-in was that the user in-

¹²<http://docs.oasis-open.org/xliff/xliff-core/xliff-core.html>

```
<xliff version="1.2">
<file original="123"
  source-language="en" target-language="fr" datatype="text"
  category="IT" product-name="TEST">
<header>
  <phase-group>
    <phase phase-name="mt_baseline" process-name="Machine Translation"
      tool="sb" tool-id="mydemoMT"
      date="2013-05-03T14:02:40.470235" contact-email="11850@test.com"/>
    <phase phase-name="start_pe" process-name="bilingual"
      tool="ACCEPT Portal" tool-id="ACCEPT Post Edit Plug-in 1.0"
      date="2013-05-03T13:17:22" contact-email=""/>
    <phase phase-name="r1.1" date="6/6/2013 1:11:47 PM"
      contact-email="test@test.com"/>
  </phase-group>
  <count-group name="1">
    <count phase-name="r1.1" count-type="x-keys" unit="instance">1</count>
    <count phase-name="r1.1" count-type="x-delete-keys" unit="instance">1</count>
    <count phase-name="r1.1" count-type="x-editing-time" unit="x-seconds">2.922</count>
    <count phase-name="r1.1" count-type="x-typing-time" unit="x-seconds">2.922</count>
  </count-group>
  <count-group name="2"/>
</header>
<body>
  <trans-unit id="1">
    <source>This is a new test</source>
    <target phase-name="r1.1">C'est un nouveau test</target>
    <alt-trans phase-name="mt_baseline">C'est une nouvelle test</alt-trans>
  </trans-unit>
  <trans-unit id="2">
    <source>the old one was very old</source>
    <target phase-name="mt_baseline">L'ancien a été très vieux</target>
  </trans-unit>
</body>
</file>
</xliff>
```

Figure 1: Report in XLIFF format

terface should be as simple as possible, using few but well-known button icons. One of the first questions we had to answer when creating the plug-in was how the text should be displayed to users. It was felt that displaying text in a grid format would be intimidating to non-professional translators. Instead, we decided to present the whole target text to edit in a column (left), thus giving users the ability to navigate from one sentence to another by clicking on individual sentences. The editing takes place in a separate column (right), segment by segment, as shown in Figure 2:

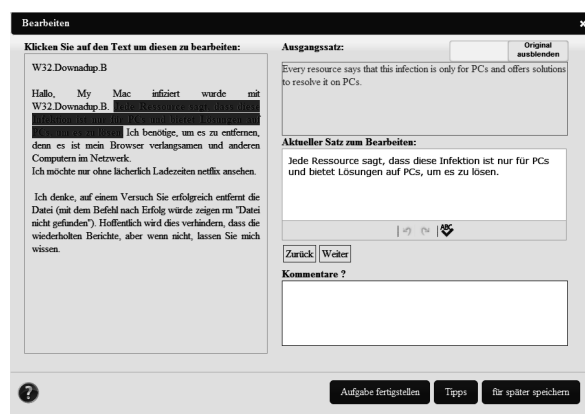


Figure 2: Client application

Figure 2 shows that the ACCEPT plug-in is target text-driven since the objective of the task is to post-edit pre-translated content. Users are invited to make edits to an existing target text rather than creating a target text from scratch by getting some

Description	XLIFF count Type	Unit	Level	Example
Total time (active editing window)	x-total-time (in phase-name="complete_pe")	Seconds	Task	
Total editing time (all revision-level editing times)	x-editing-time (in phase-name="complete_pe")	Seconds	Task	
Number of keys pressed	x-keys	Instance	Revision	
Number of space keys pressed	x-white-keys	Instance	Revision	
Number of alphanumerical and non-white keys pressed	x-nonwhite-keys	Instance	Revision	a,J,{
Number of arrow keys pressed	x-arrow-keys	Instance	Revision	↑, ←, <i>End</i> , <i>Home</i>
Number of other keys pressed	x-other-keys	Instance	Revision	CTRL,Shift
Number of triggered checks	x-checking-usage	Instance	Revision	
Selection of alternative translation options	x-trans-options-usage	Instance	Revision	
Time spent editing	x-editing-time	Seconds	Revision	
Description	XLIFF Mapping	Type	Level	Example
Used Translation Options	note annotates="target" from="trans_options"	String	Revision	Hat Ist 0
Generic Comment	note annotates="general" from="user"	String	Revision	Easy to edit!
Custom Comment	note annotates="target" from="user"	String	Revision	Terminology
Text entered	target of alt-trans if not final or target of trans-unit if final	String	Revision	Ist jemandem das schon mal begegnet?
Global Comment	note annotates="general"	String	Task	Easy!

Table 1: XLIFF elements used to display PE actions

inspiration from a source text. This choice is motivated by the fact that target users of this application are members of a community who may have a very limited knowledge of the source language. To fully support this use case, the user interface of the plug-in has already been fully localised from English into French and German, and more languages will be easily added in the future.

4.1 Displaying alternative translation options

The best output of an MT system (especially the output of an SMT system) corresponds to a product of choices between multiple options, but some of these options sometimes turn out to be sub-optimal in a given context. While experienced post-editors can easily correct these poor choices by selecting more suitable morphological, lexical or syntactic options without being prompted, it has been shown that lesser-equipped translators can benefit from having access to alternative translation options (Koehn, 2010). As described in section 3,

the JSON format used to upload data into a given project may contain such alternative options. The ACCEPT plug-in currently expects each alternative translation option to be present in an array associated with each target token. Once a token has been identified by a user (by hovering over it and clicking it), its alternative options are displayed in a contextual list, where each item can be selected in order to replace the original token in the target text. The use of this functionality is recorded in the XLIFF report using a *count* element with a *count-type* attribute whose value is *x-trans-options-usage*. The options that were actually selected by the user are recorded in *note* elements.

4.2 Checking Target Content

The output of MT systems is sometimes ungrammatical, so users may benefit from some assistance to identify those parts of the output that should be modified. Spelling and grammar checking is there-

fore made available by embedding a pre-editing plug-in in the post-editing plug-in itself. This functionality can be triggered by clicking the *ABC* icon. Once this icon is clicked, a pop-up window appears and misspelt words or ungrammatical phrases are underlined. The user can then select a suggestion or ignore the recommendation provided by the tool. Supported languages currently include English, French and German. It is currently not clear how useful the default rules are for checking MT output, but they can also be used once the MT output has been post-edited. The use of this functionality is recorded in the XLIFF report using a *count* element with a *count-type* attribute whose value is *x-checking-usage*.

4.3 Showing the source

Post-editing is traditionally believed to be most successful in a bilingual manner (i.e. post-editing with reference to the source text) for the reason that meaning that may have been lost/distorted in the machine translation process can be retrieved from the source text. While research in monolingual post-editing is scarce, especially in regards to domain experts as post-editors rather than linguists/translators, providing the post-editor with the opportunity of choosing the post-editing set-up dynamically (i.e. monolingual/bilingual) has been identified as a potential way of minimising or preventing user frustration. This is supported by feedback that has been gathered in internal studies, which indicated that users were eager to see the source, as further described in Section 5.

To illustrate switching between bilingual and monolingual modes, consider what happens if the project default is the monolingual mode. The source will then not be shown in the interface when a task is opened initially. The user can then decide to switch to being shown the original segment for the current segment. Regardless of how many switches are performed per segment, the last state the switch is in is retained for the next segment a user chooses to edit. It can be switched at any time. When the editor is closed, the page is refreshed or a new task is selected, the project default is displayed again (in this case the source is not shown).

5 Evaluation

A pilot study with 8 participants was recently conducted with a view to analyse the types of ed-

its made by volunteer post-editors (most of whom were forum users with no formal translation or post-editing experience). During this study, we found that the interface was not straight-forward to use because it differed from editors users were familiar with (e.g. MS Word). This resulted in users copying and pasting content into other editors or editing all content in one segment. In addition, the help button was not visible enough and the instructions were not clear enough. These issues were addressed by making the help button more visible and by creating a short training video for future users, consisting of a screen recording and a voice-over in the participants' native language. Since these improvements have been made, a new study has been conducted with 18 participants (five of whom had translation experience), and it has revealed that there is no longer any confusion on how to use the interface. We also took this opportunity to ask users three specific questions about the interface:

- Which feature did you like best?
- Which feature did you like least?
- Which feature did you miss most?

The results presented in Table 2 show that most users were satisfied with the feature set of the ACCEPT environment. When asked to identify the feature they liked least, half of the respondents answered "None", suggesting that everything was working as expected. Based on the feedback received, areas to improve include a more comfortable display of the target text (without any scrolling) and the ability to have access to the whole source text.

The lukewarm feedback received in relation to the spell-checking feature confirms the need to use a tool that has been specifically designed with machine translation output and post-editing in mind. Traditional spelling and grammar checkers are usually not trained on machine translation output, so they can generate false alarms when used in a post-editing context. Some of the enhancement suggestions also reveal that experienced users (e.g. people with translation expertise) are interested in having access to features found in existing translation environments (such as advanced editing tools or online dictionaries). It is also worth highlighting two other user suggestions: having to ability to

Best	%	Worst	%	Missed	%
Show source	77.8	None	50	None	50
All	11.2	Spell checker	27.8	Dictionary/thesaurus/alternative translations	10
Yellow text highlighting	5.5	“Next” button	11.2	Show whole source text	5
Whole text on left side	5.5	Small fonts	5.5	No scrolling	5
		Vanishing tasks	5.5	Show how others have translated this	5
				Revert to MT segment	5
				Use proofreading symbols	5
				Show editing tool as menu bar	5
				Display statistics and badges	5
				Have another window for draft sentences	5

Table 2: Usability survey results

revert to the original MT output and being able to see how other users may have post-edited the same segment.

6 Conclusions and future work

This paper has presented a simple, yet powerful, novel post-editing environment aimed at community users who may have very limited translation expertise. We described the architectural design choices that were made to create a flexible environment that clearly segregates the client application from the rest of the environment. User actions and interactions that take place in the client application are captured via the ACCEPT API, stored in a database, and made available in real-time to project administrators via a downloadable report based on the XLIFF format. Future work will focus on documenting the API of the ACCEPT environment, so that post-editing tasks can be created easily, without necessarily having to upload input files manually. An extension of this work will also include necessary changes to allow the use of the client application outside of the ACCEPT portal. We have already made some progress in this area by allowing certain users to make use of the ACCEPT client application inside Amazon Mechanical Turk’ HITs.¹³ We will also investigate the possibility to integrate functionality provided by advanced (S)MT systems, such as the mining of complete search graphs to display useful alternative translation options. Finally, we would also like

to conduct an evaluation comparing the usage of this environment with the usage of another existing tool to benchmark how long post-editing takes in each environment.

Acknowledgements

The work presented in this paper is being supported by the European Commission’s Seventh Framework Programme (Grant 288769). The authors would like to thank Fred Hollowood and Jason Rickard for their insights during the initial design phase, all of the users of the ACCEPT post-editing environment who have provided some feedback to date, as well as the reviewers of this paper for their comments.

References

- Wilker Aziz, Sheila Castilho, and Lucia Specia. PET: a Tool for Post-editing and Assessing Machine Translation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *LREC*, pages 3982–3987. European Language Resources Association (ELRA), 2012.
- Michael Carl. Translog-II: a Program for Recording User Activity Data for Empirical Reading and Writing Research. In *LREC*, pages 4108–4112, 2012.
- Giselle de Almeida and Sharon O’Brien. Analysing Post-Editing Performance: Cor-

¹³<https://www.mturk.com>

- relations with Years of Translation Experience. In François Yvon and Viggo Hansen, editors, *EAMT 2010, 14th Annual Conference of the European Association for Machine Translation*, Saint-Raphaël, France, 2010.
- Michael Denkowski and Alon Lavie. TransCenter: Web-Based Translation Research Suite. In *AMTA Workshop on Post-Editing Technology and Practice Demo Session*, San Diego, CA, 2012.
- Jakob Elming and Ragnar Bonk. The CAS-MACAT workbench: a tool for investigating the integration of technology in translation. In *Proceedings of the International Workshop on Expertise in Translation and Post-editing - Research and Application*, Copenhagen, Denmark, 2012.
- Raymond Flournoy and Christine Duran. Machine Translation and Document Localization Production at Adobe: From Pilot to Production. In *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, 2009.
- Spence Green, Jeffrey Heer, and Christopher D Manning. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 439–448, New York, NY, USA, 2013. ACM.
- Ana Guerberof Arenas. *Productivity and quality in the post-editing of outputs from translation memories and machine translation*. PhD thesis, Universitat Rovira i Virgili, Spain, 2012.
- Philipp Koehn. Enabling Monolingual Translators: Post-Editing vs. Options. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 537–545, Los Angeles, California, 2010.
- A Kumaran, K Saravanan, and Sandor Maurice. wikiBABEL: community creation of multilingual data. In *Proceedings of the 4th International Symposium on Wikis*, WikiSym '08, pages 14:1—14:11, New York, NY, USA, 2008. ACM.
- Linda Mitchell and Johann Roturier. Evaluation of Machine-Translated User Generated Content : A pilot study based on User Ratings. In *Proceedings of EAMT 2012*, pages 61–64, Trento, Italy, 2012.
- Mirko Plitt and François Masselot. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *Prague Bull. Math. Linguistics*, 93:7–16, 2010.
- Johann Roturier and Anthony Bensadoun. Evaluation of MT Systems to Translate User Generated Content. In *Proceedings of the Thirteenth Machine Translation Summit*, pages 244–251, Xiamen, China, 2011.