2014 AMTA
Association for Machine Translation in the Americas

Third Workshop on Post-Editing Technology and Practice

Sharon O'Brien
Michel Simard
Lucia Specia

WORKSHOP

The 11th Conference of the Association for Machine Translation in the Americas

Vancouver, BC
October 22-26
amta2014.amtaweb.org

The 11th Conference of the Association for Machine Translation in the Americas

October 22 – 26, 2014 -- Vancouver, BC Canada

Proceedings of the

Third Workshop on Post-Editing Technology and Practice

(WPTP-3)

Sharon O'Brien, Michel Simard and Lucia Specia (Eds.)

Association for Machine Translation in the Americas
http://www.amtaweb.org

Table of Contents

Proceedings of the Third Workshop on Post-editing Technology and Practice (WPTP-3)

Edited by Sharon O'Brien, Michel Simard and Lucia Specia

AMTA Workshop, Vancouver, Canada, October 26 2014

Committees

Organizing Committee

Sharon O'Brien - Dublin City University

Michel Simard - National Research Council Canada

Lucia Specia - University of Sheffield

Joss Moorkens - Dublin City University

Program Committee

Nora Aranberri -- University of the Basque Country

Diego Bartolome -- tauyou <language technology>

Michael Carl -- Copenhagen Business School

Francisco Casacuberta -- Universitat Politècnica de València

Stephen Doherty -- University of Western Sydney

Andreas Eisele -- European Commission

Marcello Federico -- FBK-IRST

Mikel L. Forcada -- Universitat d'Alacant

Philipp Koehn -- University of Edinburgh

Roland Kuhn -- National Research Council Canada

Isabel Lacruz -- Kent State University

Alon Lavie -- Carnegie Mellon University

Elliott Macklovitch -- Independent Consultant

Daniel Marcu -- University of Southern California

John Moran -- Transpiral Translation Services

Kristen Parton -- Columbia University

Maja Popović -- DFKI

Johann Roturier -- Symantec

Midori Tatsumi -- Independent Researcher/Lecturer

Andy Way -- CNGL / Dublin City University

Programme

9:00 - 10:30 Session 1

9:00 *MT Post-Editing into the Mother Tongue or into a Foreign Language? Spanish-English MT Output Post-Edited by Translation Trainees*

Pilar Sánchez Gijon and Olga Torres Hostench

9:30 *Comparison of Post-Editing Productivity between Professional Translators and Lay Users*

Nora Aranberri, Gorka Labaka, Arantza Diaz de Iiarraza, and Kepa Sarasola

10:00 *Monolingual Post-Editing by a Domain Expert is Highly Effective for Translation Triage*

Lane Schwartz

10:30 - 11:00 Coffee Break

11:00 - 12:30 Session 2

11:00 *Perceived vs. Measured Performance in the Post-Editing of Suggestions from Machine Translation and Translation Memories*

Carlos Teixeira

11:30 *Perception vs. Reality: Measuring Machine Translation Post-Editing Productivity*
Federico Gaspari, Antonio Toral, Sudip Kumar Naskar, and Declan Groves

12:00 *Cognitive Demand and Cognitive Effort in Post-Editing*
Isabel Lacruz, Michael Denkowski and Alon Lavie

12:30 - 14:00 Lunch Break

14:00 - 16:00 Posters and Demos

Vocabulary Accuracy of Statistical Machine Translation in the Legal Context
Jeffrey Killman

Towards Desktop-Based CAT Tool Instrumentation -- iOmegaT
John Moran, David Lewis and Christian Saam

Translation Quality in Post-Edited versus Human-Translated Segments: A Case Study
Elaine O'Curran

The TAUS Post-Editing Course & The TAUS Post-editing Productivity Tools
Attila Görög

QuEst for Estimating Post-Editing Effort
Lucia Specia

An Open-Source Desktop Post-Editing Tool
Lane Schwartz

Real-Time Adaptive Machine Translation: Cdec and TransCenter

Michael Denkowski

Post-Editing User Interface Using Visualization of a Sentence Structure

Yudai Kishimoto, Toshiaki Nakazawa, Daisuke Kawahara and Sadao Kurohashi

Kanjingo: A Mobile App for Post-Editing

Sharon O'Brien

16:00 - 17:30 Panel: What Lies Ahead for Post-editing?

Moderator: Mike Dillinger (AMTA President)

Panelists:

- Olga Beragovaya (Welocalize)
- John Moran (CNGL, TCD)
- David Rumsey (President-Elect at American Translators' Association)
- Lori Thicke (Translators Without Borders)
- Chris Wendt (Microsoft)

MT Post-editing into the mother tongue or into a foreign language? Spanish-to-English MT translation output post-edited by translation trainees

Pilar Sánchez-Gijón pilar.sanchez.gijon@uab.cat
Olga Torres-Hostench olga.torres.hostench@uab.cat
Tradumàtica Research Group, Departament of Translation, Interpreting and Eastern Studies, Universitat Autònoma de Barcelona, Bellaterra, 08193, Spain

Abstract

The aim of this study is to analyse whether translation trainees who are not native speakers of the target language are able to perform as well as those who are native speakers, and whether they achieve the expected quality in a "good enough" post-editing (PE) job. In particular the study focuses on the performance of two groups of students doing PE from Spanish into English: native English speakers and native Spanish speakers. A pilot study was set up to collect evidence to compare and contrast the two groups' performances. Trainees from both groups had been given the same training in PE and were asked to post-edit 30 sentences translated from Spanish to English. The PE output was analyzed taking into account accuracy errors (mistranslations and omissions) as well as language errors (grammatical errors and syntax errors). The results show that some native Spanish speakers corrected just as many errors as the native English speakers. Furthermore, the Spanish-speaking trainees outperformed their English-speaking counterparts when identifying mistranslations and omissions. Moreover, the performances of the best English-speaking and Spanish-speaking trainees at identifying grammar and syntax errors were very similar.

1. Introduction

> *V.(d) A translator should, as far as possible, translate into his own mother tongue or into a language of which he or she has a mastery equal to that of his or her mother tongue.*
> UNESCO Recommendation, November 22nd 1976

Since UNESCO issued its recommendation, more and more translation companies and translation faculties have been adopting this "mother-tongue principle", with excellent results. However, various authors have questioned this principle. Campbell (1998:212) argues that the "dynamics of immigration, international commerce and the postcolonial world make it inevitable that much translation is done into a second language, despite the prevailing wisdom that translators should only work into their mother tongue." Kelly (2003) defends the same arguments of necessity, and Pokorn (2005: *X*) is perhaps the most critical. The latter argues that the traditional view "according to which translators should translate only into their mother tongue in order to create linguistically and culturally-acceptable translations (…) stems from an aprioristic conviction unsupported by any scientific proof that translation into a mother tongue is ipso facto superior to translation into a non-mother tongue."

Is the same principle applicable to post-editing (PE)? Should PE also adhere blindly to this principle? Marcel Thelen, a supporter of non-native translators (2005:250), argues that the principle is too rigid and questions the UNESCO recommendation. The following extract from Thelen's book unintentionally became the starting point for the research presented in this paper.

> "Applying the mother tongue principle seems to have become a sort of quality assurance, part of a guarantee of specialisation. Sticking to the native speaker rule is, however, not necessary in many cases, especially since clients do not all require the same quality of translations depending on the envisaged purpose. (…) In addition, with the implementation of technology and different kinds of translation tools, it becomes increasingly 'easy' for non-natives speakers to produce good English through post-editing."

Is this true? Is it really so easy for non-native speakers to produce "good English"? In what sense would PE quality be affected if it were carried out by non-native speakers? This study attempts to discover whether non-native translation trainees could provide as good PE (in terms of accuracy and language) as native translation trainees. We conducted an empirical study in which a PE task from Spanish into English was carried out by two groups of subjects: non-native translation trainees and native translation trainees. The two groups were asked to post-edit several sentences from the user interface and help file of the OpenOffice software package. The aim of this study was to compare the results of the PE carried out by the two groups in terms of accuracy and language, and thus determine whether non-native translation trainees are able to meet the expected quality standards.

Traditionally, it was supposed there were two levels of PE —light PE and full PE— although TAUS prefers to talk about "good enough quality" and "quality similar to a human translator". In our study, we expected non-native translation trainees to achieve "good enough quality" (TAUS 2010). The above list shows that in PE that is considered "good enough quality" expectations of the quality of language used are low, whereas accuracy is very important. Accuracy is non-negotiable both in light and full PE. So, if non-native speakers are able to provide accurate PE then we would need to bring into question the mother-tongue principle for PE.

2. Related work

Native translation professionals seem to be the best suited for any PE job. Guerberof (2008, 2009, 2012) analysed the productivity and the quality of PE from the translation memories (TM) and machine translation (MT) output of native professional translators; Plitt and Masselot (2010) tested productivity by comparing MT+PE with traditional translation by native professional translators; Almeida and O'Brien (2010) compared PE performance with professional translation experience, and Temizoz's (2013) compared the differences in PE performance between engineers and professional translators. Other interesting studies of different post-editor profiles are Koehn's (2010) on PE by monolingual users and Mitchell, Roturier and O'Brien's (2013) who compare PE by monolingual users vs. that of bilingual users.

Many companies and organizations also rely on native speaking professional translators: the Commission of the European Communities worked with professionals on Systran PE (Wagner 1985); Sybase worked with professionals on PangeaMT PE (Bier and Herranz, 2011); and Continental Airlines worked with professionals on SDL PE (Beaton and Contreras, 2010).

Some organizations, however, are exploring other post-editor profiles. Computer Associates, for instance, is developing a PE crowdsourcing platform where any person who

knows two languages could become a post-editor and quality would be assessed by ranking the PE output (Muntés-Mulero and Paladini, 2012; Muntés-Mulero et al., 2012).

In an academic context, some researchers have carried out studies on PE using native translation students. Sutter and Depraetere (2012) analysed the relationship between PE, distance and fluency using translation trainees and O'Brien (2005) observed the correlation between PE effort and MT translatability. Especially relevant for our study is Garcia (2010) whose study on PE quality and the time taken for the task, used Chinese non-English native translation trainees, comparing their MT+PE with a translation made using a TM.

In light of the related literature, our contribution aims to explore a factor in the post-editors' profile that has been largely unexplored so far (except in the cases mentioned above): their mother tongue.

3. Method

In this paper, we will check the following hypothesis: "PE jobs performed by native translation trainees will be more accurate and linguistically correct than those performed by non-native translation trainees". In order to investigate whether this hypothesis is valid, we will try to answer the following research questions:

- To what extent is PE performed by non-native translation trainees accurate?
- To what extent is PE performed by non-native translation trainees linguistically correct?

As stated in the introduction, our main focus was to establish what level of PE accuracy and linguistic correctness non-native translation trainees can produce taking into account their presumed poorer use of the foreign language compared with that of native speaking translation trainees. Accuracy was analysed by evaluating post-edits of mistranslations and omissions; language was analysed by evaluating post-edits of grammatical and syntax errors. The results of this study may be useful when taking decisions on PE training programs.

3.1. Preparation of the corpus

The sentences to be post-edited were taken from the English-Spanish bitext of OpenOffice (Tiedemann 2009). We downloaded the TMX file for the en_GB and es languages (50.6k). The characteristics of this corpus made it a good choice for our study:

- All post-editors had computer skills, so the subject matter of the text did not pose a major challenge to them.
- All the sentences were easy to understand even though post-editors were not given the context.

We began by collecting sentences in Spanish from the English-Spanish bitext. We then back-translated the Spanish sentences into English using Google Translate. Finally, we compared the machine-translated sentences against the original English sentences and selected machine-translated sentences containing specific types of translation errors, as explained in section 3.2.

Twenty correct sentences, chosen from the corpus, were inserted into a table in two columns, Spanish on the left, English on the right, so that students could familiarise themselves with the genre, grammar and syntax of the text. Students read these sentences as a warm up task. We made sure that the sentences chosen were well translated and easy to understand without additional context.

A second table contained the sentences to be post-edited. The first column contained the original segment in Spanish, while the second and third columns showed the Spanish sentences translated into English using MT. Post-editors were asked to enter their changes in the third column of the document. The table below shows the first sentence as it appeared to the post-editors.

ES (do not change)	**EN (do not change)**	**EN (for post-editing)**
Crea un vínculo al arrastrar y colocar un objeto del Navegador en un documento.	Create a link to drag and drop an object from the Navigator into a document.	Create a link to drag and drop an object from the Navigator into a document.

Table 1. Extract from the PE test.

The sentences in this second table were chosen in line with the objectives of our research. In the example above, for instance, there is a mistranslation ("to drag and drop" is not the same as "al arrastrar") and a grammar mistake ("create" instead of "It creates").

Each sentence was independent from the others, but they were clear and comprehensible even without any context. All the English sentences given to the participants were raw back-translations from Spanish into English produced by Google Translate. These back-translations contained the errors listed in the error typology presented above: 10 mistranslations, 5 omissions, 5 grammatical errors, and 5 syntax errors.

3.2. Error typology

We defined our error typology based on the error types listed in the "List of MQM Issue Types", the QT LaunchPad project (2013), and the TAUS Error typology guidelines (as drafted by Sharon O'Brien for TAUS Labs and reviewed and endorsed by a large number of companies and organizations in 2013). The TAUS typology has four main categories: accuracy, language, terminology and style. As defined in the TAUS Error typology guidelines, "the category of 'Accuracy' is applied when incorrect meaning has been transferred or there has been an unacceptable omission or addition in the translated text." In our test, we divided the TAUS Accuracy category into two subcategories, mistranslations (when an incorrect meaning has been transferred) and omissions. We did this in order to observe which mistakes made by post-editors were due to misunderstanding the message (accuracy) or which were due to lack of sufficient attention (omissions). Many words in our MT output were misplaced, resulting in mistranslations and changes of meaning. When preparing the test, we also found omissions that affected the meaning of the target sentence. Additions were not included in the study as there were no additions in the MT output analysed. For the study we selected ten sentences containing a mistranslation and five containing an omission.

As for Language errors, we identified two subcategories: grammar and syntax. Although the two are closely related, we wanted to analyse them separately since grammatical errors may reduce comprehensibility more than syntax errors. For the study we selected 5 sentences with a grammatical error and 5 with a mistake in the syntax.

We disregarded stylistic mistakes because this is a rather subjective category and is largely irrelevant when dealing with "good enough" PE. We also disregarded terminology mistakes because they are applied only "when a glossary or other standard terminology source has not been adhered to" (TAUS, 2013) and we did not provide a separate glossary for students. Moreover, we selected more mistranslation errors (10) than errors in the other categories (5) because, in our view, this category is key in translation quality reports. Mistranslation

errors are heavily penalised and translation trainees must be aware of the dangers of mistranslation in post-editing.

3.3. Participants

Fifteen translation trainees participated in the research: 12 Spanish non-native English speakers (group A) and 3 native English-speakers from the USA (group B). The group of non-native speakers is larger than the group of native speakers because we were especially interested in the results of this group and its variability. As a matter of fact, the results for group A were initially compared with the correct solution, rather than with the results of group B. We are aware that, statistically, the analysis of the performance of these two small groups of participants will not yield solid results, but they should show if this issue is worthy of further research.

3.3.1. Group A

Of the 12 participants in group A, 11 were female and 1 male, (9 were aged 21 or 22, with the remaining 3 aged between 27 and 38). All were students in their seventh semester of a translation and interpreting degree in Spain. English was the first foreign language of 10 of the students, the second foreign language of 1 and the third foreign language of 1. Nine of the students had taken a Spanish-English translation course in Spain and 2 had done so in England via an Erasmus grant.

All members of the group attended an introductory 2-hour seminar on machine translation PE and worked on a 10-hour group project to compare and post-edit output from different MT systems into their mother tongue. Their attitude towards PE was generally positive. Only 1 participant said she was "not at all interested" in PE; 3 said they were "a little interested" and 7 said they were "quite interested". The group was fairly used to using technology. Most of them were active computer users (for professional and personal purposes) but were not computer experts (only one said she had ever developed and created new computing solutions).

3.3.2. Group B

Group B was composed of 3 native English-speaking American translation trainees aged 30-40, all of whom were female. All three were very successful students taking a distance master's programme in professional translation with grades of 95% or higher. All had taken Spanish-English and English-Spanish translation courses and had some specific training in localization. Two of them had completed professional internships. All members of the group received the same introductory seminar on machine translation and PE as Group A, part of an online course in translation technologies. Afterwards, they worked on a 5-hour individual project to post-edit a text produced by MT. Two participants said they were "very interested" or "extremely interested" in PE, while 1 participant said she was only "a little interested". Two of them used IT resources both for professional and personal purposes, while 1 stated that she did not like IT resources but had learned to use them professionally.

3.4. Post-editing test procedure

Both groups participated from the same location where they had received their translation classes, and training in MT and PE. Group A participants worked in a computer lab and group B participants worked online. Members of the two groups were given only one Microsoft Word document with all the information they needed for the test.

Firstly, students read a general explanation of the task to be carried out. They were asked to focus particularly on mistranslations, omissions, and grammar and syntax errors that compromised translation accuracy and comprehensibility. Once they had read the instructions they were invited to answer a profile questionnaire. They were then instructed to read the first table with the 20 correct bilingual sentences, and were told to pay close attention so they could become accustomed to the genre. After this warm up task, they noted the time and then started the PE. The sentences appeared in random order. After PE, they noted the time again and signed a research authorization form granting us permission to use the test data anonymously and only for research purposes.

The post-edited segments were analyzed anonymously. When processing the post-edited segments we considered only those translation errors that we used as indicators. Correctly edited segments were counted as successful edits; segments in which the error had not been detected were counted as unsuccessful edits; and sentences in which changes introduced by post-editors did not make it clear whether the error had been detected were ignored. The post-edited segments were compared with the correct and published translation of the sentences of the corpus.

4. Results

4.1. Results for non-native Spanish-to-English participants (group A)

Table 2 below shows the overall success rates for group A. The results show that non-native participants were most successful at detecting mistranslations, followed by omissions, syntax errors and finally grammatical errors.

Error category	Success rate	Error category	Success rate
Mistranslations	72.5%	Accuracy	72%
Omissions	71.67%		
Syntax	66.67%	Language	59%
Grammar	51.67%		

Table 2. Success rates for group A

It is worth mentioning that non-native participants performed much better correcting mistranslations than language errors. In other words, their command of written English is not as good as their attention to details with respect the accuracy of sentences. We will now analyse each category separately.

a) *Mistranslations*. The test contained ten previously identified mistranslations. Variability in this category is quite high. The least corrected mistranslation was corrected by 5 of the 12 participants, while one mistranslation – the final one – was corrected by all 12 participants, perhaps because by that point the participants were more familiar with the task.

b) *Omissions*. The test contained five omissions that we identified in the MT output. The least-corrected omission was corrected by only 5 of the 12 post-editors, probably because MT produced a perfectly coherent, grammatically correct, factually accurate sentence.

c) *Syntax*. The test contained five previously identified syntax errors. In this category, the least corrected error was corrected by 7 of the 12 participants. The inability to correct this syntax error might be due to a lack of knowledge of the finer rules of English grammar by the participants. The most corrected mistake was corrected by 11 of the 12 participants.

d) *Grammar*. The test contained five previously identified grammatical errors. Uncorrected grammatical errors revealed that non-native participants had some problems with English grammar. The least corrected sentence was corrected by only 5 of the 12 participants, while the most corrected sentence was corrected by 9 of them.

Results for group A have been compared in line with the time devoted to the task by each participant. Figure 2 shows the number of post-editings by each post-editor alongside the time they devoted to the test (upper line). Interestingly, in some cases, there seems to be a correlation between the time devoted to the task and the results. For instance, the best performer of the group spent more time than the rest (except for one other person who spent even more time on the task) while the worst performer was the fastest in the group.

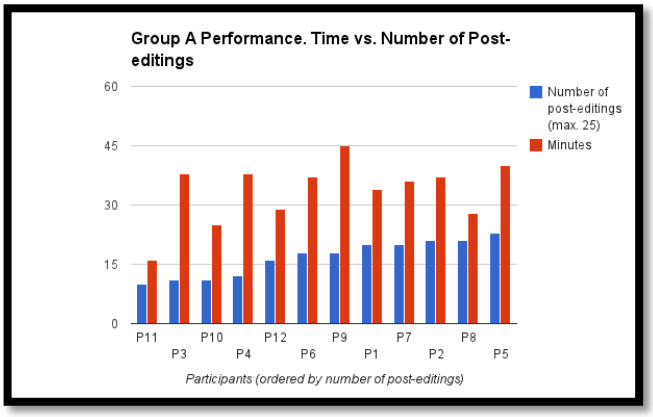


Figure 1. Number of post-editings vs. time spent - Group A.

In Figure 1. we can also see the variability of results among participants. Variability ranges from only 8 correct post-edits by the worst performer to 24 by the best performer. The mean number of correct edits per person was 16.75 out of 25.

Finally, in these results it is interesting to note the results of participants who had a slightly different background to the rest. For instance, all but two participants had English as their first foreign language. For participant 3, English was her third foreign language and she scored 13 out of 30. However, three other participants scored less despite English being their first foreign language. English was participant 9's second foreign language, and she scored 20 out of 30, outscoring five participants whose first foreign language was English. We can see, then, that participants whose first foreign language was not English were not the worst in the group. However, their results suggest they would not be trustworthy post-editors for translations into English.

If we only look at the overall results we might conclude that working with non-native speakers would not be advisable. However, good non-native post-editors could be suitable for the job. In this PE test, if only the participants having a 70%+ success rate were selected then the overall results would change dramatically, because in average their performance would be comparable to group B.

4.2. Results for Spanish-to-English post-editing carried out by native speakers (group B)

Table 3 below shows that native participants corrected almost all the syntax errors. They also performed very well at correcting mistranslations, followed by omissions and grammar. It is

worth mentioning that their results for Accuracy and Language almost match, indicating a more equal level of skills than non-native participants.

Error category	Success Rate	Error category	Success Rate
Mistranslations	80%	Accuracy	76%
Omissions	66.67%		
Syntax	93.33%	Language	77%
Grammar	60%		

Table 3. Success rates for participants in group B

From the individual questionnaires we know that participants who had completed a professional internship were slightly more successful. Let us look at the results of each error category in more detail:

a) *Mistranslations*. Overall performance by this group in this category was good; as was expected. None of the participants in this group corrected all the mistranslations, but they came very close. Two participants, both of whom had completed a professional internship, successfully corrected 90% of them.

b) *Omissions*. None of the participants corrected all the omissions, and one of the incorrect sentences was missed by all the participants.

c) *Syntax*. Two of the participants in group B corrected all 5 syntax errors, while the third missed just one of the errors. Syntax is closely related to style and English native speakers are expected to perform very well at correcting this kind of error. Our results suggest this is the key error category that makes the difference between native and non-native speakers.

d) *Grammar*. Grammatical errors were not very obvious. They were grammatical errors caused by a bad translation, but sometimes they resulted in an apparently correct sentence with a different meaning. As was the case with omissions, native speakers of English may have missed some of these errors because the target sentence did not seem to be ungrammatical. Only 1 participant detected all the grammatical errors, while 1 detected just one of the incorrect sentences.

Results for group B have been compared using the time devoted to the task by each participant. Figure 2 shows the number of post-editings by each post-editor alongside the time they devoted to the test (upper line). In this group, there is no obvious correlation between the time devoted to the task and the results of each participant.

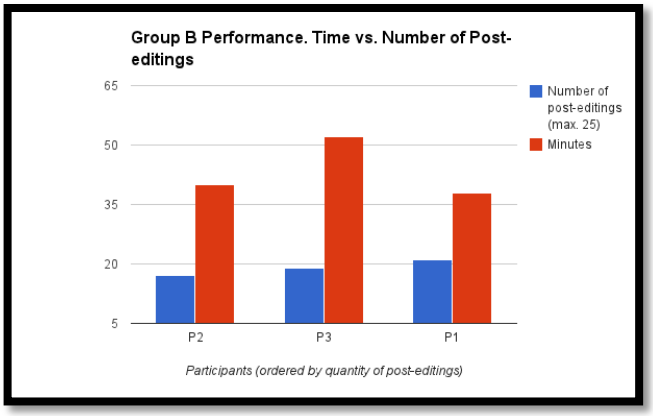


Figure 2. Number of post-editings vs. time spent: Group B

4.3. Comparison of results

As Table 4 shows, group B (native speakers) performed better than group A (non-natives) in all error categories except Omissions. The biggest difference between group A and B is found in the Syntax category, due to the better command of the English language by native participants. However, Accuracy results are more similar in both groups. Upon a closer examination it is worth observing deviation results. Needless to say that the low number of participants, especially in group B, made it difficult to extract valid conclusions about deviation, other than the fact that there was huge variability among participants.

	Group A	Group B
Mistranslations	72.5% (σ = ±21.1%)	80% (σ = ±17.32%)
Omissions	71.67% (σ = ±24.8%)	66.67% (σ = ±23.09%)
ACCURACY	72.22% (σ = ±20.26%)	75.56% (σ = ±19.24%)
Grammar	51.67% (σ = ±21.67%)	60% (σ = ±40%)
Syntax	59.17% (σ = ±27.12%)	93.33% (σ = ±11.55%)
LANGUAGE	59,17% (σ = ±19.29%)	76.67% (σ = ±15.28%)
Total	64.4% (σ = ±23.38%)	75.37% (σ = ±22.11%)

Table 4. Average success rates achieved by each group for each error category

General results seem to validate our hypothesis "PE jobs performed by native translation trainees will be better than those performed by non-native translation trainees". Nevertheless, a more detailed analysis per category, looking at each group's average scores for each category, as well as the highest and the lowest score in each group could shed light on the strengths and weaknesses of each group in order to find some answers to our research questions: "(1) To what extent is PE performed by non-native translation trainees accurate?; (2) To what extent is PE performed by non-native translation trainees linguistically correct?"

a) *Mistranslations*. Both groups achieved high scores when correcting mistranslations.

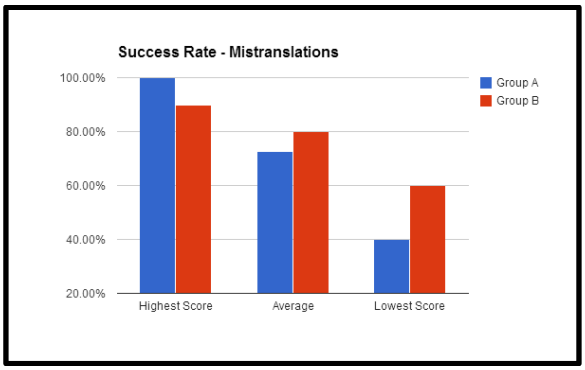


Figure 3. Success rates for each group in identifying mistranslations

Despite group B being the most successful at detecting mistranslations as a whole, it is worth noting that the best performer at correcting mistranslations belonged to group A (as well as the worst performer). Group B's results were more homogeneous. Group B participants were less prone to missing mistranslations, with the lowest scorer scoring considerably more than the lowest scorers in group A. In this case, being a native speaker or not does not seem to be essential in order to deliver a good post-editing job.

b) *Omissions*. As can be observed in Figure 4, group A was more successful at detecting omissions than group B. This might have been because many raw MT sentences were linguistically correct, but with a different meaning to the original source sentence. Since the target sentences seemed to be correct, native speakers of English may have been misled into believing they were valid translations. This is an important warning for native translation trainees doing PE.

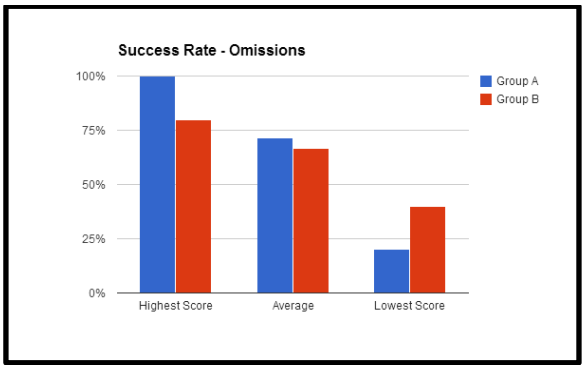


Figure 4. Success rates for each group in identifying omissions

c) *Grammar*. A priori, we may have expected all participants to have been successful at correcting grammatical errors. However, grammatical errors produced by MT systems range from the very obvious (for instance where a pronoun does not match the gender of a nearby subject) to the inconspicuous (for instance where the gender is given in a previous sentence). In other words, although the MT system might propose a sentence that is grammatically correct, the sentence might not be an appropriate equivalent of the source sentence in

terms of its grammar. Only a careful reading of the source sentence would detect this kind of problem.

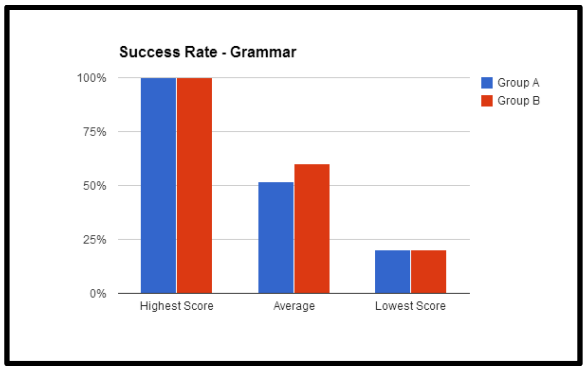


Figure 5. Success rates for each group in identifying grammatical errors

In this case, results from both groups were very similar. Detecting and correcting inconspicuous grammatical errors seems to be more related to the personal skills of each participant rather than the general skills of each group. In both groups there are participants who solved 100% of errors and only 20% of errors.

d) *Syntax*. Language command seems to make a difference in detecting syntax errors.

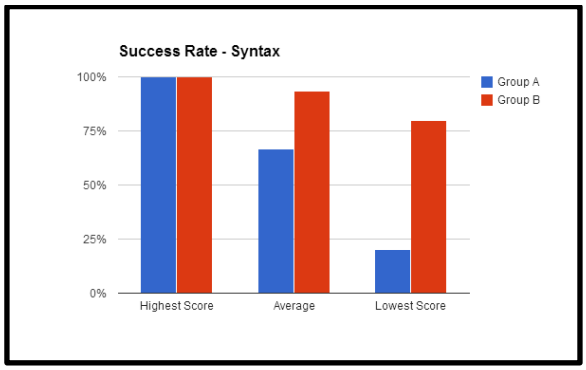


Figure 6. Success rates for each group in identifying syntax errors

As shown in Figure 6, group B performed better and more homogeneously than group A. Although the textual genre chosen for this study (text embedded in a software UI and text from the software's help module) has just a few very distinctive features (it is relatively poor stylistically), participants with less competence in the target language had more difficulty correcting errors related to word order. This suggests that native speakers of the target language would perform better than non-native speakers when post-editing texts belonging to a syntactically and stylistically more complex genre. If we look at the lowest scores from each group in this area we see that the lowest-scoring participants in group B considerably outperformed the lowest-scoring participant in group A.

So far we have presented the results for the best and the worst performer noting that that there is considerable variability between them. Now it is worth commenting on variability in individual results. If we look at the results of individual participants we find that there were very good post-editors in both groups (see Figure 7). This information is very useful when

qualifying the validity of the hypothesis as from the detailed results in Figure 7, we would argue that it cannot be automatically inferred that PE jobs performed by native translation trainees are more accurate and linguistically correct than PE jobs performed by the best non-native translation trainees. Indeed, quality seems to depend on the person, regardless of their mother tongue.

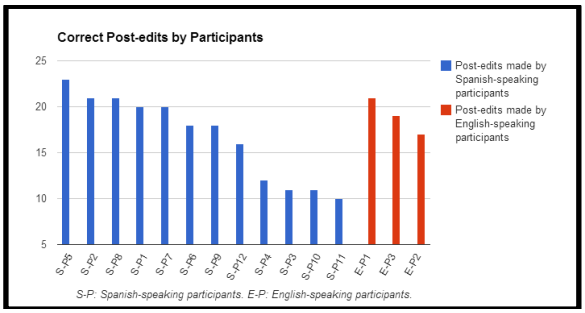


Figure 7. Correct post-edits by participants

When it comes to quality, the same principle is applicable for the research questions "(1) To what extent is PE performed by non-native translation trainees accurate? (2) To what extent is PE performed by non-native translation trainees linguistically correct?" When comparing the best Spanish-speaking participant with the best English-speaking participant (see Figure 8), it is revealing that one Spanish-speaker's post-editing job is more accurate and even more linguistically correct than that of the best English-speaking participant.

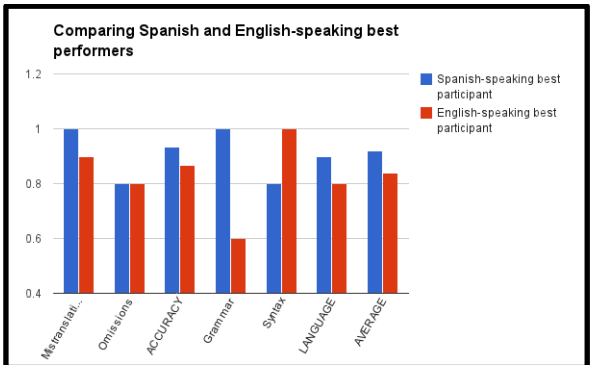


Figure 8. Correct post-edits by participants

5. Conclusion and future work

Our study suggests that good performers who are not native speakers of the target language can do "good enough" PE jobs from Spanish to English. The results for the best participants from group A (non-native) were very similar to those of group B (native) in a "good enough" PE task. These results question the validity of the "mother-tongue principle" for "good-enough" PE, as there are better and worse participants regardless of their mother tongue. The results also show that, while not all non-native participants were suitable for PE tasks, good non-native participants can indeed be suitable.

In light of the results, we need to ask whether non-native translation trainees require a different kind of PE training to that offered to native speaking translation trainees. These

results suggest that non-native translation trainees need more practice in identifying syntax and grammatical errors, while native translation trainees need to develop greater observational skills and pay more attention to detail in order to detect inconspicuous errors in natural sounding, grammatically correct sentences. From a cognitive point of view, it would be useful to prepare training material designed to develop observational and error detection skills for PE. These skills have proved to be more decisive in detecting accuracy errors in PE than the established mother-tongue principle. Besides providing an argument in favour of the creation of exercises related to PE guidelines, error typology, etc., these results suggest that PE training should also include exercises oriented towards the development of observational and error detection skills. In future work, it would be interesting to identify what skills good native and non-native post-editors have in common so that these may be improved through training.

Finally, it could be very useful for the translation industry to set up a standardized test to help identify which translators perform better as post-editors and whether they are best at post-editing into their mother tongue or into a foreign language. The test could be adapted for different PE projects or clients, with different types of errors. Such a test would be a suitable way of evaluating both native and non-native post-editors' skills, instead of merely disregarding non-natives a priori. In fact, such a test may even completely disregard if candidates post-edit into their mother tongue.

Acknowledgement

This research is supported by the *Red de formación de poseditores* (Post-editor Training Network) (Ref. FFI2011-16021-E), funded by the Spanish State Secretariat of Research, Development and Innovation), and by the research project *Accesibilidad lingüística y sensorial: tecnologías para las voces superpuestas y la audiodescripción* (Linguistic and sensory accessibility: technologies for voiceover and audio description) (Ref. 2012-31024), funded by the Spanish State Secretariat of Research, Development and Innovation).

References

Beaton, A. and Contreras, G. (2010) Sharing the Continental Airlines and SDL post-editing experience. Proceedings, Ninth Conference of AMTA, Denver, Colorado.

Bier, K. and Herranz, M. (2011) MT Experiences at Sybase. Presentation, Localization World Barcelona, July 2011. http://www.slideshare.net/manuelherranz/loc-world2011-kbiermherranz-8730502. Accessed 3 January 2014.

Campbell, S. (1998) Translation into the second language. Addison Wesley Longman, New York.

De Almeida, G. and O'Brien, S. (2010) Analysing post-editing performance: correlations with years of translation experience. Proceedings, 14th Annual Conference of the EAMT, S. Rafael, France.

De Sutter, N., Depraetere, I. (2012) Post-edited translation quality, edit distance and fluency scores: report on a case study. Proceedings, Journée d'études "Traduction et qualité Méthodologies en matière d'assurance qualité", Université Lille 3, Sciences humaines et sociales, Lille.

García, I. (2011) Translating by post-editing: is it the way forward? Mach Transl, 25:217–237.

Guerberof, A. (2009) Productivity and quality in MT post-editing. Proceedings, MT Summit XII workshop: beyond translation memories: new tools for translators, Ottawa, Canada.

Guerberof, A. (2012) Productivity and quality in the post-editing of outputs from translation memories and machine translation. Doctoral thesis, Universitat Rovira i Virgili, Tarragona.

Guzmán, R. (2007) Manual MT Post-editing: if it's not broken, don't fix it. Trans J, 11, http://www.bokorlang.com/journal/42mt.htm. Accessed 24 January 2014.

Kelly, D. (2003) La direccionalidad en la traducción e interpretación: perspectivas teóricas, profesionales y didácticas. Atrio, Granada.

Koehn, P. (2010) Enabling Monolingual Translators: Post-Editing vs. Options. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 537–545.

Proceedings, AMTA Workshop on MT Research and the Translation Industry, pp 21-31.

Enabling Monolingual Translators: Post-Editing vs. Options

Roturier, M. and O'Brien, S. (2013) Community-based post-editing of machine-translated content: monolingual vs. bilingual. Proceedings, MT Summit XIV Workshop on Post-editing Technology and Practice, Nice, pp 35-43.

Muntés, V., Paladini, P. (2012) Bringing the crowd in the post-editing process. Presentation, TAUS European Summit, Paris, http://www.youtube.com/watch?v=1bgIKPULcsc. Accessed 3 January 2014.

Muntés-Mulero, V., Paladini, P., Solé, M. and Manzoor, J. (2012) Multiplying the Potential of Crowdsourcing with Machine Translation. Proceedings, Tenth Biennial Conference AMTA, San Diego.

OPUS, the open parallel corpus website, http://opus.lingfil.uu.se/. Accessed 3 January 2014.

O'Brien, S. (2005) Methodologies for measuring the correlations between post-editing effort and machine translatability. Mach Transl, 19:37–58

O'Brien, S. (2009) Researching and teaching post-editing. Presentation, Seminar Post-Editing MT Output: Views from the researcher, trainer, publisher and practitioner, CNGL, http://www.mt-archive.info/MTS-2009-OBrien-ppt.pdf. Accessed 3 January 2014

Plitt, M., Masselot, F. (2010) A productivity test of statistical machine translation post-editing in a typical localisation context. Prague Bull Math Linguist, 93:7–16.

Pokorn, NK (2005) Challenging the traditional axioms: translation into a non-mother tongue, John Benjamins,Philadelphia, PA / Amsterdam.

QT Launch Pad (2013) http://www.qt21.eu/launchpad/content/list-mqm-issue-types. Accessed 3 January 2014

TAUS (2013) Quality evaluation using an error typology approach. https://evaluation.taus.net/resources/error-typology-guidelines. Accessed 3 January 2014

TAUS (2010) Machine translation post-editing guidelines. https://evaluation.taus.net/resources/guidelines/post-editing/machine-translation-post-editing-guidelines. Accessed 3 January 2014

Temizoz, O. (2013) Postediting machine translation output and its revision: Subject-Matter Experts versus Professional Translators. Doctoral thesis, Universitat Rovira i Virgili, Tarragona.

Thelen, M. (2005) Translating into English as a non-native language: the Dutch connection. In: Anderman G, Rogers M (eds) In and out of English: for better, for worse, Multilingual Matters, Clevedon, pp 242-255.

Tiedemann, J. (2009) News from OPUS – A collection of multilingual parallel corpora with tools and interfaces. In: Nicolov N, Mitkov R (eds.) Recent advances in natural language processing (vol V), John Benjamins, Amsterdam/Philadelphia, pp 237-248.

Wagner, E. (1985) Rapid post-editing of Systran. In: Lawson V (ed) Tools for the trade: translating and the computer 5. Aslib, London, pp 199-213.

Comparison of post-editing productivity between professional translators and lay users

Nora Aranberri nora.aranberri@ehu.es
Gorka Labaka gorka.labaka@ehu.es
Arantza Diaz de Ilarraza a.diazdeilarraza@ehu.es
Kepa Sarasola kepa.sarasola@ehu.es
IXA Group, Department of Computer Languages and Systems,
University of the Basque Country, Donostia, 20018, Spain

Abstract

This work compares the post-editing productivity of professional translators and lay users. We integrate an English to Basque MT system within Bologna Translation Service, an end-to-end translation management platform, and perform a producitivity experiment in a real working environment. Six translators and six lay users translate or post-edit two texts from English into Basque. Results suggest that overall, post-editing increases translation throughput for both translators and users, although the latter seem to benefit more from the MT output. We observe that translators and users perceive MT differently. Additionally, a preliminary analysis seems to suggest that familiarity with the domain, source text complexity and MT quality might affect potential productivity gain.

1. Introduction

Thanks to the significant improvement of machine translation (MT) over the past two decades, the translation industry has already started to exploit it, mainly by combining it with post-editing. A good number of recent works report a productivity increase thanks to post-editing of MT output as compared to the traditional human translation (e.g., Guerberof, 2009; Plitt and Masselot, 2010; Garcia, 2011; Pouliquen et al., 2011; Skadiņš et al., 2011; den Bogaert and Sutter, 2013; Green et al., 2013; Läubli et al., 2013).

Most post-editing research is designed with professional translators in mind (even if often post-editors involved in experiments are non-professionals and students). However, it is not only language professionals who can benefit from MT in their daily tasks but also regular users who might need to perform a translation sporadically. In this work, we aim to compare the post-editing productivity between professional translators and lay users. We consider the regular example of professional translators working for a language service provider (LSP) and the particular context of administrative and staff members at the University of the Basque Country. This institution is set in a bilingual cultural context. All legal documentation and administrative communication must be provided in Spanish and Basque and study programmes are offered in both languages in parallel. In this scenario, university employees often find themselves having to produce the same material in two languages, that is, having to translate. This work aims to examine the potential benefit of MT during translation for these users as well as for professionals.

Post-editing research has so far focused on mainstream languages. An added challenge of this work is the use of an English to Basque MT system for post-editing. Research on Basque MT has been ongoing for a few years now (Diaz de Ilarraza et al, 2000a; 2000b; Labaka

et al., 2007; España-Bonet et al., 2011; Mayor et al., 2011). However, Basque being a low-resourced language, researchers and developers have found themselves with limited resources to build competitive MT systems and automated translation has not been included within the translation processes of local LSPs yet. To our knowledge, this is the first (open) productivity experiment done for the English to Basque translation direction.

Laurenzi et al. (2013) pointed out the existence of many communities that could benefit greatly from machine translation but, as in the case presented in this work, have not yet started to use it, as authors suggest, either due to lack of awareness or barriers to adoption. The work in Laurenzi et al. (2013) presents a feasibility study to introduce MT coupled with post-editing in local and regional health departments in the United States. It highlights a number of requirements the translation platform should address, such as being intuitive and easy to install, allowing users to share ongoing and completed jobs. Our work builds on this first feasibility study and goes a step forward by assessing the actual translation performance. We identify a suitable tool for our users that is intuitive, easy to access and allows sharing translation resources such as translation memories (TM) or specialized MT engines and measure productivity gain, while comparing it with the performance of professional translators.

2. Experimental Design

In the following sections we describe the platform and the texts used during the experiment, we present the profile of the participants and detail how the productivity test was set up.

2.1. The Platform

The post-editing environment used in the experiment was the Bologna Translation Service (BTS), the product of an EU-funded ICT PSP 4th Call, Theme 6: Multilingual Web project (ID 270915).[1] It is an end-to-end web-based translation management tool in which users with different roles (manager, requester, reviewer, etc.) participate on-line at different stages of the translation workflow. It couples translation memory (TM) and machine translation (MT) capabilities within a simple work environment. BTS was designed with lay users in mind. The work environment offers a simple layout with a top bar with the main action buttons and job information (see Figure 1). Below, the source text is split into segments and the target side is filled with either TM (fuzzy-)matches or MT candidates for the reviewer to work on. It is a plain tool as opposed to more sophisticated software developed in the CASCAMAT[2] and MateCat[3] projects (Alabau et al. 2013; Federico et al., 2012), which include interactive translation prediction and track post-editing operations.

[1] http://www.bologna-translation.eu/
[2] http://www.casmacat.eu/
[3] http://www.matecat.com/

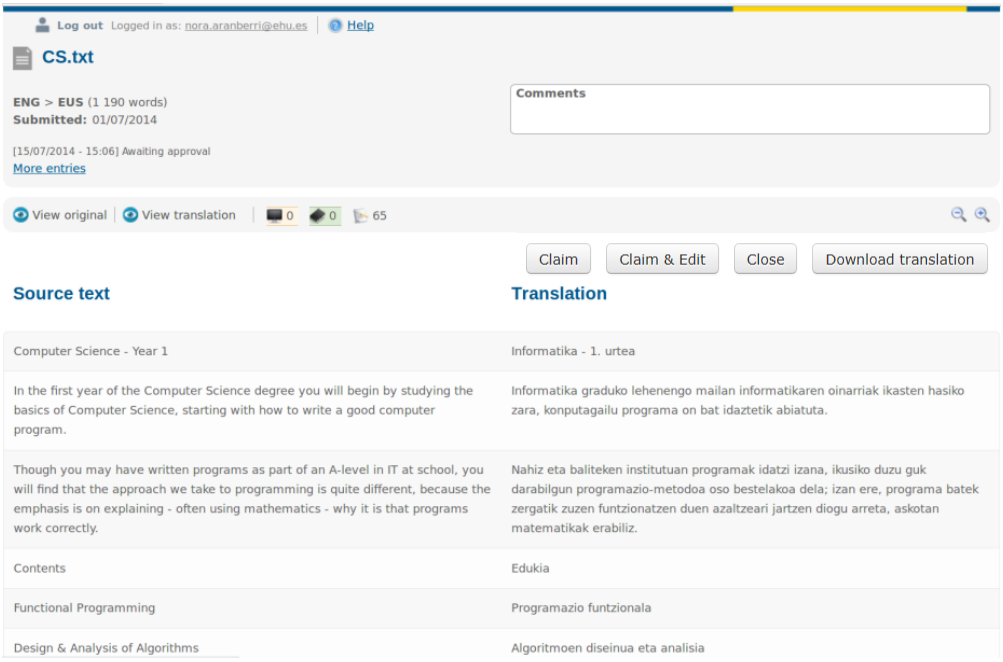


Figure 1. Screenshot of the translation environment at BTS.

For the current experiment, the BTS platform was enhanced with an English to Basque MT system. A standard phrase-based statistical machine translation system was built based on Moses using a parallel corpus of 14.58 million English tokens and 12.50 million Basque tokens (1.3 million parallel sentences) which includes localization texts (graphic user interface strings and user documention), academic books and web entertainment data. To address the token mismatch between English (analytic language) and Basque (agglutinative language) tokens, the aligner was fed with segmented words for the agglutinative language. Several segmentation options exist: we can isolate each morpheme, or break each word into lemma and a bag of suffixes; we can establish hand-written rules for segmentation, or let an automatic tool define and process the words unsupervised. Based on the results from Labaka (2010), we opted for the second option and joined together all the suffixes attached to a particular lemma in one separate token. Thus, on splitting a word, we generated, at most, three tokens (prefixes, lemma and suffixes). Moses was trained and optimized on segmented text. Note that when using segmented text for training, the output of the system is also segmented text. Real words are not available to the statistical decoder. This means that a generation postprocess (unsegmentation step) is needed to obtain real word forms. We incorporated a second language model (LM) based on real word forms to be used after the morphological postprocess. We implemented the word form-based LM by using an n-best list, as was done in Oflazer and El-Kahlout (2007). We first asked Moses to generate a translation candidate ranking based on the segmented training explained above. Next, these candidates were postprocessed. We then recalculated the total cost of each candidate by including the cost assigned by the new word form-based LM in the models used during decoding. Finally, the candidate list was re-ranked according to this new total cost. This somehow revises the candidate list to promote the ones that are more likely to be real word-form sequences. The weight for the word form-based LM was optimized at Minimum Error Rate Training (Och, 2003) together with the weights for the rest of the models.

2.2. The Texts

Two texts of around 1200 words each were selected for the experiment. Because the SMT system was trained on science and localization texts among others, it was deemed convenient to use texts from related domains. Text A consists of short 1st year computer science course descriptions from a UK university, and Text B is a collection of six short science articles from www.sciencenews.org, the flagship website by the Society for Science & the Public (SSP), dedicated to public engagement in scientific research and education.

The selection of the texts was somehow also motivated by the prospective profile of the lay user group. Whereas translators are professionals who are trained to handle a large variety of topics, we aimed to engage staff members of the Faculty of Computer Science as lay users. Therefore, we considered that they would feel more comfortable dealing with texts from computing and scientific domains.

Both texts are very similar from a size point of view. Text A consists of 1190 words and 65 sentences, whereas Text B consists of 1196 words and 67 sentences. Moreover, a wide variety of sentence-lengths are present in the texts, ranging from 1 to 51 words in length. Similarly, both texts display a moderate degree of difficulty, as they address specialized topics. In particular, terminology is very significant in both texts. Text-types, in contrast, are different. Text A is mainly descriptive and Text B, although descriptive, tends to be more literary. The literariness might leave some room for creativity in the transations, but at the same time, this might pose extra difficulty, particularly for lay users.

2.3. The Participants

We aim to compare the productivity gain for professional translators and for lay users. To this end, the same experiment was conducted with a group of translators and a group of users. The former consisted of six professional translators who regularly work for the Elhuyar Foundation (see Table 1).[4] They reported a translation experience ranging between 4 and 18 years. Four out of six had never performed post-editing before, whereas two reported having participated in previous MT experiments. They completed the translations required by the experiment as if they were regular jobs with the difference that they used the BTS platform instead of the usual TM tool (SDL Trados).

	T1	T 2	T 3	T 4	T 5	T 6
Translation experience	8 years	12 years	4 years	18 years	20–23 years	8 years
PE experience	experiments	no	no	no	experiments	no

Table 1. Translation and post-editing experience of professional translators.

The lay user group was represented by five lecturers and one post-doc from the Faculty of Computer Science at the University of the Basque Country (see Table 2). They report a level of English ranging from B2 to C1, and their level of Basque ranges from C1 to C2. As mentioned below, the University of the Basque Country is set in a bilingual cultural context and study programs are offered in both Spanish and Basque in parallel (with English becoming more and more a third language of instruction). In this scenario, some lecturers have "bilingual" job positions, which means that they might find themselves teaching the same modules in two languages. As a result, they need to prepare the same study material in both languages. The participants in the lay user group report having little translation experience, and when any, this seems to be mainly from Spanish to Basque or viceversa. They report never

[4] http://www.elhuyar.org/EN

using MT engines for this purpose. In fact, they admit to most often re-writing the material rather than translating it sentence by sentence.

	U1	U2	U3	U4	U5	U6
Level of English	B2	B2	B2	C1	C1	C1
Level of Basque	C1	C1	C1	C2	C1	C2
Translation experience	little Spanish-Basque, no MT	little no MT used	little no MT used	little no MT used	no	re-write of material

Table 2. Translation experience and language proficiency of lay users.

2.4. The Productivity Test

We aimed to measure changes in productivity (if any) by monitoring the difference in the time spent translating the texts with and without the help of our MT system's output. In order to do so, BTS was programmed to count the time participants needed to complete the job. As described in section 2.1, source texts are provided to participants segmented on the left column, and a translation box per segment is opened for him/her to work on on the right column. More precisely, therefore, BTS would record the time the participants spent in each segment by saving the time each translation box was active. This method also opened the possibility for participants to work on a single segment multiple times if necessary. We asked participants to avoid distractions when completing the translation/post-editing job so that extra time would not count towards the time spent on the job. Although we recommended keeping them at a minimum, the possibilities to pause the job, and log in and out of the platform were provided to ensure that saved times were as accurate as possible.

Each participant worked under both setups, with and without the help of the MT output. In each group, texts were assigned to each participant in a way that each text was translated and post-edited three times, and the same text was not assigned to the same participant for both setups.

Simple guidelines were provided to participants with information about how to use the platform, as well as the job they should complete. They were given total freedom as to the resources they could use to perform the job (dictionaries, web searches), the only restriction being that they should not use an external MT engine. Therefore, once participants registered in the BTS platform and were assigned the two tasks, they could decide freely when and where to complete the jobs. They were given 1–2 weeks to complete the tasks. All the previous conditions should help simulate a real translation scenario as much as possible for both professional translators and lay users.

The BTS platform includes a translation memory (TM) feature. During this experiment, however, translators and users would work with an empty TM so that we only focus on the difference in translation throughput (average number of words translated per hour) and no other parameters such as fuzzy-match repetition rates introduce noise in the data.[5] We assume that the use of TMs would affect both setups (translation from scratch and post-editing) equally, and therefore, no effort was made in compiling TMs for this occasion. Nonetheless, the TM feature was activated so that the translations were stored segment-aligned and can serve as parallel data for future NLP-related tasks.

[5] The current version of the BTS does not include fuzzy-match propagation of the translations validated within the project.

3. Results

3.1. Professional Translators

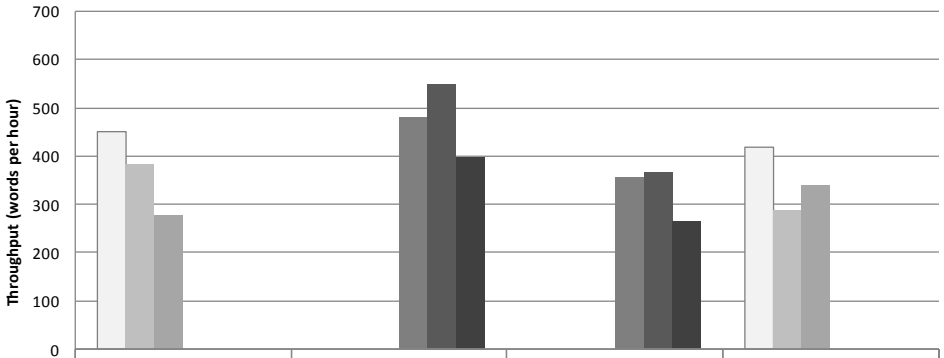
By looking at the average throughput per setup, we see that overall, post-editing our MT system obtains a slightly higher throughput than translating from scratch (17.66%). The experiment shows that, on average, the throughput increased for both texts with the aid of MT, although at different levels. The throughput for Text A increased from 372 words per hour to 477 words per hour (28.22%) and for Text B from 330 words per hour to 350 words per hour (6.06%).

If we look at the performance of individual translators, we observe that T1 and T2 have their throughput lowered with the introduction of MT 7.30% and 24.54%, respectively. However, T3, T4, T5 and T6 increased their throughput 35.01%, 49.59%, 21.43%, and 50.37%, respectively. Table 3 presents the post-editing productivity ratio (PPR), the ratio of the post-editing speed to translation speed (both expressed in words per hour). Figures less than 1 indicate cases where post-editing decreased translation throughput. Figures above 1, in contrast, indicate the ratio in which post-editing increased translation throughput.

	T1	T2	T3	T4	T5	T6
Post-editing productivity ratio	0.93	0.75	1.35	1.49	1.21	1.50

Table 3. Post-editing productivity ratio for professional translators.

If we consider closely the combination of setup, translator and text, we see that it is the three translators who post-edited Text A that benefited from the introduction of MT, as well as the slowest translator for Text A, who post-edited Text B (see Figure 2).



	Translation TEXT A	Post-editing TEXT A	Translation TEXT B	Post-editing TEXT B
T1	452			419
T2	383			289
T5	280			340
T3		482	357	
T4		549	367	
T6		400	266	

Figure 2. Throughput per text, setup and translator.

3.2. Lay User Group

The average post-editing throughput for the lay user group (434 words per hour) also surpassed the average translation throughput (386 words per hour) in 12.43%. Reinforcing the trend observed with translators, post-editing for Text A increases productivity 45.13% whereas post-editing Text B does not (-19.44%).

The performance of individual users shows that U1, U3, U4 and U5 increase translation productivity when using MT output. U2 and U6, in contrast, do not (see Table 4 for PPRs).

	U1	U2	U3	U4	U5	U6
Post-editing productivity ratio	1.69	0.63	1.76	1.02	1.08	0.89

Table 4. Post-editing productivity ratio of lay users.

Once again, if we look closely to the combination of setup, user and text, we observe that the users who benefited most from the use of MT were those who post-edited Text A. U4, who barely increased productivity, post-edited Text B, similarly to U2 and U6, who saw their throughput lowered when post-editing.

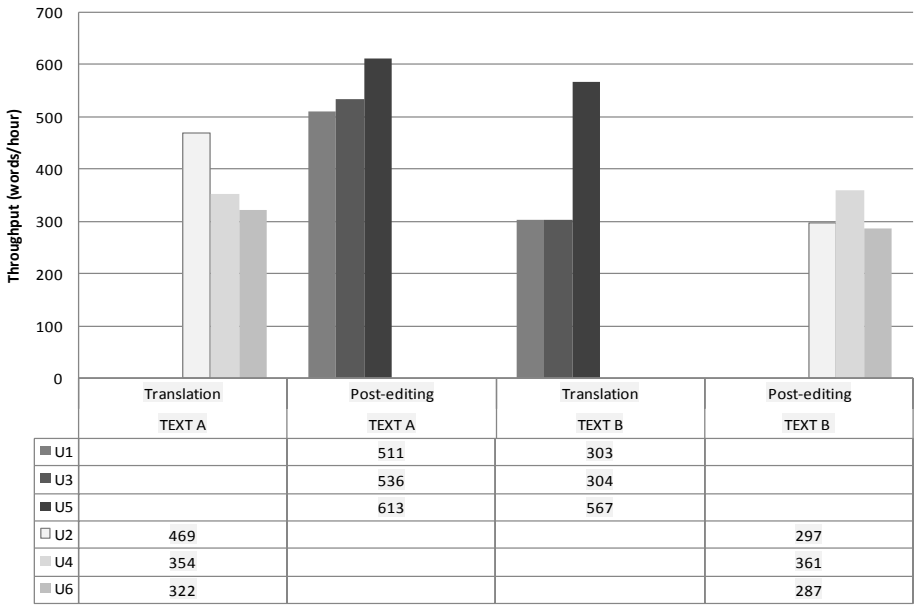


Figure 3. Throughput per text, setup and user.

4. Discussion

Results show that overall, both professional translators and lay users benefited from the use of our MT engine during translation. Translators have obtained an overall productivity gain of 17.66% and users 12.43%. However, the gain seems to be dependent on the text participants worked on. Translators have benefited more when post-editing Text A (28.22% increase) as opposed to Text B (6.06% increase). Users show an increased productivity of 45.13% when post-editing Text A, but post-editing Text B slows down the job in about 19.44%. The latter is

mainly the result of U5's high translation rate for Text B, almost double that of U1 and U3, and the low post-editing performance for Text B.

If we compare the performance of individual participants, we see that four translators and four users improve their throughput when post-editing, whereas two translators and two user do not (these last four post-edited Text B). Most participants benefit from the use of MT but not all.

Given these results, we briefly consider three factors that emerged during the experiment and might shed some light on the outcome of the experiment itself and hint to features that a company might want to consider when exploiting the MT system: attitude towards post-editing and training, source text difficulty and MT quality. Although they are all intertwined, we will consider them separately here.

4.1. Post-editing Skills and Attitude

Firstly, the skill and willingness of translators and users to make use of the MT output might affect the translation process. Post-editing has been claimed to be a different task from that of translating and one that requires different cognitive abilities and practical skills (Krings and Koby, 2001; O'Brien, 2002). Translators, therefore, need to be trained to maximize the potential benefit of post-editing. Several industrial players nowadays offer post-editing courses.[6] These usually provide an overview of MT development for the users to get acquainted with the intricacies of the systems so that they learn to interpret the output, and tips about features and patterns to watch out for in order to maximize the reuse of the output. Similarly, translator training centres have started to introduce machine translation and post-editing content within their curricula.[7] Participants in our experiment had no experience or training in post-editing. This does not allow measuring the maximum post-editing benefit.

The attitude towards post-editing or, more generally, MT might also set the tone for the job and highlight the importance of more objective measurements rather than basing the integration of MT on translator perception only. The mismatch between the translators' perception of productivity and their actual productivity has been previously reported by Autodesk, specifically on the company's follow up work on Plitt and Masselot (2010).[8] To check for this effect, we asked participants to fill in a short questionnaire after completing the tasks. One of the questions asked was whether they thought having the MT output helped them complete the translation. They were asked to mark this on a scale from -5 to 5 where -5 meant that the MT output had greatly hindered their work and 5 meant that the MT output had greatly helped their work (see Table 5). Overall, users were more positive about having the MT output displayed when translating, with only one out of six claiming that it hindered the process. In the case of translators, however, three out of six heavily penalized its use, one reported that it was better than not having it, and two reported some benefit. T1 commented that translation and post-editing required different skills and that should the same time be spent in post-editing and translating, the translation would most probably be of better quality. T6 was the most positive of all with regards to MT and admitted that the output helped in acquiring the terminology but was hopeless with syntax, which needed a complete rework. T2, T3 and

[6] For examples of training courses see TAUS's online course in collaboration with Welocalize at https://evaluation.taus.net/post-editing-course-pricing or SDL's course at http://www.translationzone.com/learning/training/post-editing-machine-translation/index-tab2.html#tabs.

[7] See an example at: MSc in Translation Technology. Dublin City University. http://www.dcu.ie/prospective/deginfo.php?classname=MTT#

[8] See http://langtech.autodesk.com/productivity.html for results of a 2-day translation and post-editing productivity test with 37 participants that Autodesk held in August of 2011.

T4 indicated that the MT output had clearly interfered in their job. T2 reported that MT output slowed down the process considerably because reading, understanding and considering what to reuse from it was very time-consuming. T3 commented that translating from scratch was easier and faster, and that even checking the MT output for terminology would most often not help. T4 claimed that the MT system did not translate the order of the phrases properly, which rendered the translation incomprehensible. Interestingly, T3 and T4 did benefit from post-editing.

U1, U2 and U3 reported that the terminology and certain chunks suggested by the MT system were useful, even when they claim to have reworked the sentences completely. U4 argued that given her lack of familiarity with the domain (Text B), she found it difficult to decide whether the terminology proposed by the MT system was correct, and therefore, she would still look up the terminology in an external source. She commented that MT output could have been a potential benefit should she be familiar with the domain of the text to be translated.

	T1	T2	T3	T4	T5	T6
PE help	0	-5	-4	-5	2	1
	U1	U2	U3	U4	U5	U6
PE help	3	2	2	0	-2	3

Table 5. Perception of MT output help during translation ([-5,5] range, where -5 greatly hinders translation and 5 greatly helps translation).

4.2. Difficulty of Source Texts

Secondly, the difficulty of the texts also needs to be considered. We will highlight two aspects here. Firstly, the familiarity with the domain of the texts; secondly, the linguistic complexity. Professional translators stated they were not familiar with the domains covered by the texts. This probably means that they were not used to the terminology and phraseology of the given domains. In contrast, lay users were a post-doc and lecturers of computer science, the domain covered by Text A. This aspect seems to be reflected in the participants' perception towards text difficulty.

As part of the questionnaire, participants were asked to specify the difficulty of the texts in a scale of 1–5, where 1 was very easy and 5 was very difficult (see Table 6). Translators reported the texts slightly differing in difficulty, but they did not agree which, Text A or Text B, was more difficult. T1 argued that Text A was *more specialized* whereas T5 considered Text B *more difficult to understand and very technical*. T2 found that Text A was easier to translate because *the whole text followed the same thread*. In contrast, T3 commented that Text A was *very abstract and disjointed*, whereas Text B was *believable and interesting*. Users, on the other hand, show a clearer tendency with three out of six identifying Text A as easy to moderate and Text B as difficult to very difficult. All users explicitly commented on their familiarity with Text A.

Questions	T1	T2	T3	T4	T5	T6
Text A difficulty	4	3	3	4	4	4
Text B difficulty	3	5	4	3	3	5
	U1	U2	U3	U4	U5	U6
Text A difficulty	2	4	4	3	2	5
Text B difficulty	4	4	4	5	5	3

Table 6. Perception of text difficulty (1–5 range, where 1 is very easy and 5 is very difficult).

Even when we are aware that many other factors are involved in the process, we turned to a readability – reading difficulty – measurement as a proxy for translation difficulty. We calculated the number of hard words[9], lexical density[10] and Gunning Fog Index (Gunning, 1952) (see Table 7).[11] When comparing both texts, we see that Text A has a slightly lower number of hard words, 11.94% as opposed to 13.64% for Text B. Lexical density is considerably higher for Text B, which means that repetitions are lower. Given that both Text A and Text B have very similar number of words, we conclude, therefore, that Text B has a higher number of different words, making it more complex. Finally, the Fog Index confirms that more years of education are necessary to read Text B. Overall, readability features suggest that Text B is more difficult to read.

	Text A	Text B
Hard Words	142 (11.94%)	162 (13.64%)
Lexical Density	34.57%	53.87%
Fog Index	11.88	12.34

Table 7. Readability-related measurements for Text A and Text B.

Understanding the source text is a vital step in translation, but other factors such as the mapping of the concepts and grammatical features pay an important role. In an attempt to measure both sides of the translation process in terms of complexity, we have also analysed the linguistic complexity of the translations produced by the participants. Based on the linguistic analysis presented in Gonzalez-Dios, et al. (2014), we have calculated the average occurrences of a number of linguistic features (including lexical, morphological, syntactic and pragmatic features) present in the translations and post-edited versions of Texts A and B (see Table 8). We observe that out of the 96 features studied, Text B has a higher number of occurrences for 63 for both translators and users, and Text A for 25 and 17, for translators and users, respectively, having no occurrences for 8 and 6 features. Additioanlly, we have considered the 10 most predictive features for complexity according to the same authors, which include a number of the most predictive features according to Feng et al. (2010), namely, part-of-speech ratios for nouns. We see that Text B appears to be more complex, scoring higher in 7 out of the 10 features.

	Text A Translators	Text A Users	Text B Translators	Text B Users
Number of features analysed	96	96	96	96
Number of features with more hits[12]	25	17	63	63
Number of features with no hits	8	6	8	13
Ratios				
Proper nouns / common nouns	0.01077	0.01830	**0.15715**	0.15672
Appositions / noun phrases	0.04433	0.03040	0.13129	**0.14345**
Appositions / all phrases	0	0.00065	**0.00293**	0.00331
Named entities / common nouns	**0.14422**	0.18222	0.11357	0.11184
Unique lemmas / all lemmas	0.03586	0.01747	0.05376	**0.06150**
Acronyms / all words	**0.24811**	0	0.21422	0.03030
Causative verbs / all verbs	**0.00137**	0.00035	0.00061	0.00030

[9] For readability testing hard words are those with three or more syllables.
[10] Lexical density is provided as the type/token ratio x 100.
[11] The Gunning Fox Index returns the number of years of education that a reader hypothetically needs to understand a particular text. It is calculated by multiplying by 0.4 the sum of the average number of words in the sentences and the percentage of hard words. For instance, the New York Times has an average Fog Index of 11-12.
[12] Counts normalized for 1000 words.

Modal-temporal clauses / subordinate clauses	0	0	**0.00047**	0.00000
Destinative case endings / all case endings	0	0	**0.00085**	0
Connectors of clarification / all connectors	0.16157	0.14564	0.25437	**0.25869**

Table 8. Comparison of linguistic complexity features for Text A and Text B translations and post-edited versions.

A final aspect that is worth noting is text expansion rates. English is an analytic language and Basque is an agglutinative language, which usually means that word-counts contract when translating into Basque. For translators, on average, Text A has contracted to 90.25% and Text B has expanded to 103.85% with respect to the English source. For users, both texts contract but whereas Text A goes down to 85.21%, Text B still remains at a high 98.21%. The fact that an expansion has occurred in Text B might be due to participants tending to over-explain or paraphrase. This might be a result of the complexity of the content.

4.3. MT Quality

MT output quality is also essential in post-editing measurements, a factor that is often neglected when reporting productivity gain. An exception is a seminal work by Koehn and Germann (2014). They studied the relation between MT quality and post-editing, and concluded that differences in post-editing skills might be more decisive than MT quality to foresee productivity gain when comparing systems within the same quality range. Our findings show that the text for which a higher increase in productivity was obtained seems to be slightly easier and better suited for our MT system.

In order to test for MT quality, we calculated BLEU and TER scores on Text A and Text B using the translations and post-editings obtained during the experiment as references (see Table 9). If we consider the scores obtained with the translations as references, we see that Text B obtains a slightly higher BLEU score than Text A, but output for Text A is better according to TER. If we take post-editings or all six versions as references, then Text B seems to get a higher quality output. As expected, we observe that the post-edited versions obtain significantly better BLEU and TER scores as the post-edited versions resemble the MT output more than translations made from scratch. Overall, automatic score results would lead us to conclude that the MT output for Text B might be slightly better, and therefore more reusable.

Translators	Translations		Post-editings		All 6 references	
	BLEU	TER	BLEU	TER	BLEU	TER
Text A	12.33	68.07	23.31	55.70	25.26	55.44
Text B	12.71	71.26	27.26	55.32	28.61	54.07
Users	**Translations**		**Post-editings**		**All 6 references**	
	BLEU	TER	BLEU	TER	BLEU	TER
Text A	8.44	77.78	39.65	52.32	41.00	51.27
Text B	10.45	79.35	29.17	56.41	30.49	56.44

Table 9. BLEU and TER scores for Texts A and B using participants's texts as references.

It is worth noting the difference in BLEU scores between the post-edited versions of translators and users. A higher BLEU score means that more of the MT output was kept in the final versions. Users, therefore, accepted or reused a considerably larger amount of MT output than translators. A possible interpretation is that the MT output was of a relatively good quality and users, domain experts of Text A, were easily able to identify reusable chuncks and exploit terminology much more so than translators.

Finally, to test whether the MT engine was better prepared to address Text A or Text B, we calculated perplexity and out-of-vocabulary (OOV) words. Perplexity is used as a mea-

surement of how well the language model predicts the reference translations. The smaller the perplexity, the more and longer overlap exists between the reference and the language model in the MT system. This measurement shows that the MT engine is better suited to output a correct version for Text A than for Text B (see Table 10). Note that the high perplexity values, calculated per word, are in line with those reported for morphologically rich languages (see Popel and Mareček, 2010).

	Translators			Users	
	Text A	Text B		Text A	Text B
Translation 1	988.220	1274.940	Translation 1	2086.65	1359.14
Translation 2	898.022	1081.580	Translation 2	1423.39	1794.06
Translation 3	776.408	966.309	Translation 3	1686.82	1944.15
Post-editing 1	850.781	909.031	Post-editing 1	842.705	1341.37
Post-editing 2	660.740	909.031	Post-editing 2	1002.280	1300.49
Post-editing 3	688.585	984.311	Post-editing 3	902.018	1173.91

Table 10. Perplexity calculated on 5-grams.

In Table 11 we see the number of OOVs in the training corpus with respect to the source texts. Once again, Text A seems better suited for our MT system, as only 0.7% of the tokens were missing from the training data, as opposed to the 4.9% for Text B. This is yet another feature that hints that MT output for Text A might be of better quality than for Text B.

	Training sentences	Training tokens	Training types	Tokens	Types	OOV tokens	OOV types
Text A	1,290,501	15,798,942	221,172	1292	420	9 (0.7%)	8 (1.9%)
Text B	1,290,501	15,798,942	221,172	1381	645	68 (4.9%)	32 (5.0%)

Table 11. OOV counts for the Text A and B together with information on training data.

5. Conclusions

We have integrated an English to Basque MT system within BTS, an end-to-end translation management platform, and performed a post-editing producitivity experiment in a real working environment to compare the performance of professional translators and (prospective) lay users of BTS. Results suggest that overall post-editing increases translation productivity for both translators and users, although the latter seem to benefit more from the MT output specially when working on their domain of expertise.

We have observed that translators and users perceive MT output differently. Overall, translators seem to find that it interferes and slows down their work. However, users do not show a negative attitude towards it and profit more from it, specially when working on a familiar domain. Although we addressed them separately, we saw that textual complexity and MT quality are connected and seem to affect potential productivity gain. We observed that, although both texts under study were considerably specialised, the text that had higher readability and less linguistic complexity, and that was better fitted for our MT engine obtained a larger increase in productivity gain.

Acknowledgements

We would like to thank the staff members from the Faculty of Computer Science who willingly agreed to participate in the experiment.

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007/2013) under REA grant agreement nº 302038.

References

Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., González, J., Koehn, P., Leiva, L., Mesa-Lao, B., Ortiz, D., Saint-Amand, H., Sanchís, G. and Tsoukala, C. (2013). CASMACAT: An open source workbench for advanced computer aided translation. *The Prague Bulletin of Mathematical Linguistics*, 100:101–112.

Diaz de Ilarraza A., Mayor A., Sarasola K. (2000a). Building a Lexicon for an English-Basque Machine translation System from Heteogeneous Wide-Coverage dictionaries. In *Proceedings of MT 2000: machine translation and multilingual applications in the new millennium*, University of Exeter, United Kingdom: 19-22 November 2000, pages 2.1–2.9.

Diaz de Ilarraza A., Mayor A., Sarasola K. (2000b). Reusability of Wide-Coverage Linguistic Resources in the Construction of a Multilingual Machine Translation System. In *Proceedings of MT 2000: machine translation and multilingual applications in the new millennium*, University of Exeter, United Kingdom: 19-22 November 2000, pages 16.1–16.8.

España-Bonet, C., Labaka, G., Diaz de Ilarraza, A., Màrquez, L. and Sarasola, K. (2011). Hybrid Machine Translation Guided by a Rule–Based System. In *Proceedings of the Thirteenth Machine Translation Summit (MT Summit XIII)*, Xiamen, China, pages 554–561.

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)*.

Feng, L., Huenerfauth, M., Jansche, M. and Elhadad, N. (2010). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Coputational Linguistics (COLING): Posters*, pages 276–284, Beijing, China.

García, I. (2011). Translating by post-editing: is it the way forward? *Machine Translation*, 25(3):217–237.

Gonzalez-Dios, I., Aranzabe, M., Diaz de Ilarraza, A. and Salaberri, H. (2014). Simple or Complex? Assessing the readability of Basque texts. In *Proceedings of the 5th International Conference on Computational Linguistics (COLING)*, Dublin, Ireland, pages 334–344.

Green, S., Heer, J. and Manning, C. (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 Workshop on New Tools for Translators.

Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill: New York.

Koehn, P. and Germann, U. (2014). The impact of machine translation quality on human post-editing. In *Workshop on Humans and Computer-assissted Translation*, pages 38–46, Gothenburg, Sweeden. Association for Computational Linguistics.

Krings, H. and Koby, G. (eds) (2001). *Repairing Texts: Empirical Investigations of Machine-Translation Post-Editing Processes*. Kent State University Press: Kent, Ohio.

Labaka, G. (2010). EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language. Its use in SMT-RBMT-EBMT hybridation. PhD Thesis. University of the Basque Country.

Labaka, G., Stroppa, N., Way, A. and Sarasola, K. (2007). Comparing Rule-Based and Data-Driven Approaches to Spanish-to-Basque Machine Translation. In Proceedings of MT-Summit XI, Copenhagen, pages 297–304.

Mayor, A., Alegria, I., Diaz de Ilarraza, A., Labaka, G., Lersundi, M. and Sarasola, K. (2011). Matxin, an open-source rule-based machine translation system for Basque. Machine Translation Journal, 25(1), pages 53–82.

O'Brien, S. (2002). Teaching post-editing: a proposal for course content. In Proceedings of the Sixth EAMT Workshop Teaching Machine Translation, pages 99–106, Manchester, UK.

Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan.

Oflazer, K. and El-Kahlout, I. D. (2007). Exploring Different Representation Units in English-to-Turkish Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 25–32, Prague, Czech Republic.

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.

Popel, M. and Mareček, D. (2010). Perplexity of n-Gram and Dependency Language Models. In P.Sojka et al. (Eds.): TSD 2010, LNAI 6231, pages 173–180.

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 *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT*), pages 5–12.

Skadins, R., Purins, M., Skadina, I. and Vasiljevs, A. (2011). Evaluation of SMT in localization to under-resourced inflected language. In *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, pages 35–40.

Van den Bogaert, J. and De Sutter, N. (2013). Productivity or quality? Let's do both. In Proceedings of the Machine Translation Summit XIV, pages 381–390.

Monolingual Post-Editing by a Domain Expert is Highly Effective for Translation Triage

Lane Schwartz lanes@illinois.edu
Department of Linguistics, University of Illinois at Urbana-Champaign, Urbana IL, USA

Abstract

Various small-scale pilot studies have found that for at least some documents, monolingual target language speakers may be able to successfully post-edit machine translations. We begin by analyzing previously published post-editing data to ascertain the effect, if any, of original source language on post-editing quality. Schwartz et al. (2014) hypothesized that post-editing success may be more pronounced when the monolingual post-editors are experts in the domain of the translated documents. This work tests that hypothesis by asking a domain expert to post-edit machine translations of a French scientific article (Besacier, 2014) into English. We find that the monolingual domain expert post-editor was able to successfully post-edit 86.7% of the sentences without requesting assistance from a bilingual post-editor. We evaluate the post-edited sentences according to a bilingual adequacy metric, and find that 96.5% of those sentences post-edited by only a monolingual post-editor are judged to be completely correct. These results confirm that a monolingual domain expert can successfully triage the post-editing effort, substantially reducing the workload on the bilingual post-editor by only sending the most challenging sentences to the bilingual post-editor.

1 Introduction

Post-editing is the process whereby a human user corrects the output of a machine translation system. The use of basic post-editing tools by bilingual human translators has been shown to yield substantial increases in terms of productivity (Plitt and Masselot, 2010) as well as improvements in translation quality (Green et al., 2013) when compared to bilingual human translators working without assistance from machine translation and post-editing tools. More sophisticated interactive interfaces (Langlais et al., 2000; Barrachina et al., 2009; Koehn, 2009b; Denkowski and Lavie, 2012) may also provide benefit (Koehn, 2009a).

Small-scale studies have suggested that monolingual human post-editors, working without knowledge of the source language, can also improve the quality of machine translation output (Callison-Burch, 2005; Koehn, 2010; Mitchell et al., 2013), especially if well-designed tools provide automated linguistic analysis of source sentences (Albrecht et al., 2009). Schwartz et al. (2014) confirmed this result with eight monolingual post-editors on a larger 3000 sentence test corpus.

Using a bilingual judge, we evaluate the post-edited test English sentences using the 10-point adequacy metric (see Table 5) of Albrecht et al. (2009). The results of our evaluation indicate that over 95% of post-edited sentences are completely correct translations that adequately convey the meaning of the respective French source sentence. Our bilingual judge estimated that approximately 15 minutes of total effort would be required for a bilingual French-English speaker to correct the remaining 5% of post-edited sentences.

2 Effects of Original Source Language on Post-Editing Quality

In discussing post-editing, there may be cases where shared task evaluation data may have an unintended effect on post-editing quality. When a shared task test set for a particular language pair (for example, from Russian into English) is created, some portion of that shared test set may have originally been written in the shared task target language, and then professionally translated into the shared task source language. By examining the data from the 2014 Workshop on Statistical Machine Translation, we have confirmed that this is indeed the case for (at least) the Russian-English shared task.

Schwartz et al. (2014) performed a post-editing experiment as part of the WMT 2014 Russian-English shared task. The post-editors in that study anecdotally reported an effect on post-editing difficulty based on original source language: Schwartz et al. noted:

> Interestingly, several post-editors self-reported that they could tell which documents were originally written in English and were subsequently translated into Russian, and which were originally written in Russian, based on observations that sentences from the latter were substantially more difficult to post-edit. Once per-document source language data is released by WMT14 organizers, we intend to examine translation quality on a per-document basis and test whether post-editors did indeed perform worse on documents which originated in Russian.

This effect, if it does indeed exist, could mean that positive post-editing results such as those reported by Schwartz et al. (2014) may be artificially high, due to the presence of sentences in the test set which were originally written in English. Such sentences may have maintained the original English word order even after translation through Russian, and so may have been easier to translate than sentences originally authored in Russian, which might be expected to be more difficult due to more idiomatic Russian word order.

Before exploring our own post-editing study in Section 3, we therefore find it useful to conduct some further data analysis on previously released data to attempt to ascertain what effect, if any, the original source language may play in post-editing quality. After the workshop, the WMT 2014 organizers released information regarding the original source language of each sentence in the shared task test sets. In addition, as part of their WMT 2014 submission, Schwartz et al. (2014) made available the post-edited translations from their Russian-English submission, along with the results of their manual evaluation.

Schwartz et al. (2014) report that their machine translations were post-edited by a group of eight individuals. We divide their post-edited translations by original source language and by post-editor, along with the binary adequacy judgements reported for each post-edited translation. Table 1 presents the results of this data collation. For each monolingual post-editor, the percentage of sentences judged to be correct according to a monolingual human judge are broken down according to the language in which test documents were originally authored. For 7 out of 8 post-editors, we observe worse translation quality for sentences originally authored in Russian when compared to sentences originally authored in English. The overall percentage of sentences judged to be correct, taken across all post-editors, is 14 percentage points lower for sentences originally authored in Russian (57% correct) when compared to sentences originally authored in English (71% correct). Interestingly, we see no coherent effect when quality is measured using BLEU (see Table 2); for some post-editors, BLEU scores are higher (more positive) for sentences originally authored in English, but for most post-editors, BLEU scores for some post-editors are higher (in some cases by more than 5 BLEU points) for sentences originally authored in Russian.

These partially contradictory results could be an artifact of metrics, or indicative of other factors at play. In Section 3, we examine one factor that may play a more important role in

	Post-Editor								
	1	2	3	4	5	6	7	8	All
en % correct	78%	67%	78%	62%	67%	48%	64%	72%	71%
ru % correct	65%	69%	52%	51%	63%	40%	60%	43%	57%

Table 1: For each monolingual post-editor in Schwartz et al. (2014), the percentage of sentences judged to be correct according to a monolingual human judge, broken down according to the language in which test documents were originally authored.

	Post-Editor								
	1	2	3	4	5	6	7	8	All
English	27.97	21.08	25.20	28.16	27.94	21.22	23.34	24.10	25.56
Russian	27.38	26.82	24.21	27.18	28.98	22.73	28.92	26.03	26.62
difference	0.59	-5.74	0.99	0.98	-1.04	-1.51	-5.58	-1.93	-1.06

Table 2: Analysis of post-edited translation data from Schwartz et al. (2014), showing case-sensitive BLEU scores per post-editor broken down according to the language in which test documents were originally authored.

predicting post-editing quality: domain expertise.

3 Monolingual Post-editing by a Domain Expert

It has been proposed (Schwartz et al., 2014) that post-editing machine translations may be more successful when the post-editor is highly familiar with the subject matter being translated. In this section we test that hypothesis by asking a domain expert to post-edit machine translations of a French scientific article (Besacier, 2014) into English.

We begin by copying the headers, content sentences, and other text comprising Besacier (2014) from the original PDF document into plain text format (UTF-8 encoding), dividing the text into 241 distinct segments.[1] To better facilitate machine translation, each segment was placed on its own line.

The plain text of the French source document was translated using Google Translate (Google, 2014), Systran Server 7.4.2 (Systran, 2010), and Moses (Koehn et al., 2007). Google Translate is a proprietary online statistical translation system that makes use of phrase-based translation methods. Systran Server is a proprietary translation system that is primarily rule-based, although recent versions allow for hybrid rule-based/statistical functionality; we did not make use of hybrid functionality in this experiment. Moses is the de-facto standard open source phrase-based statistical machine translation system. In our experiments, Moses was trained and tuned on French-English data from IWSLT 2013, following the procedures described in Kazi et al. (2013).

The monolingual post-editor is a native speaker of English with no training or experience in French with domain expertise in the scientific article being post-edited. For each sentence, the monolingual post-editor was presented with the machine translation results produced by the three aforementioned machine translation systems. The post-editor was free to choose any of the three MT output segments as the starting point for post-editing, and was free to incorporate portions of any or all of the three MT output segments into the final post-edited result. No

[1] While most of the segments are sentences, some segments are section headers, table elements, footnotes, etc. Throughout we will use the terms segment and sentence interchangeably.

Confident	The monolingual post-editor is confident that the post-edited translation conveys the meaning of the French sentence
Verify	The monolingual post-editor believes that the post-edited translation conveys the meaning of the French sentence, but the translation should be verified by a bilingual post-editor
Partially unsure	The monolingual post-editor is not confident that a specific portion of the post-edited translation is correct; that section should be handled by a bilingual post-editor
Completely unsure	The entire sentence should be handled by a bilingual post-editor

Table 3: Confidence guidelines for monolingual post-editors.

	Post-Editor Confidence			
	Completely unsure	Partially unsure	Verify	Confident
# sentences	8	13	11	209
% sentences	3.3%	5.4%	4.6%	86.7%

Table 4: Post-editor confidence in the adequacy of post-edited translations. Confidence labels are defined in Table 3.

interactive post-editing software was provided to the post-editor; for each sentence, the post-editor was presented with the three MT output segments, and was instructed to type a fluent English output sentence into a text editor.

For each segment, the monolingual post-editor was instructed to record confidence according to the guidelines shown in in Table 3. Post-edited segments marked as "Verify" or "Partially unsure" were passed on to a bilingual post-editor to verify and correct, if necessary. Post-edited segments marked as "Completely unsure" were passed to a bilingual post-editor to post-edit or translate from scratch. Table 4 shows the breakdown of post-edited sentences by post-editor confidence. We observe that the monolingual domain expert post-edited 86.7% of the sentences without requesting assistance from a bilingual post-editor.

In determining confidence in a post-edited segment, we expect the monolingual post-editor to consider the segment's coherence with surrounding segments, and its semantic consistency with the entire document, taking into account the post-editor's own expertise in the domain. Because the monolingual post-editor does not know the source language, there is no guarantee that post-edited segments in which the post-editor is confident completely and correctly convey the meaning present in the respective source segments. For this reason, in Section 4 we perform a bilingual adequacy evaluation over all post-edited segments.

4 Evaluation

A high rate of post-editor confidence (as seen in Table 4) is worthy of note only if the post-editor's confidence is justified by corresponding high quality in post-edited results. Most machine translation experiments report quality according to BLEU or some other automated metric, as judged against one or more reference translations. In our case, the results of our work represent the only known translation of the document in question — as such, no reference trans-

lation is available.

4.1 Post-editor Confidence and Translation Adequacy

To determine the quality of post-edited translations, we asked a bilingual judge to rank the adequacy of the post-edited translations. The judge is a native English speaker fluent in French who was not involved in translating or post-editing any segments in this task. The bilingual judge was asked to rate the adequacy of all post-edited segments, using the evaluation guidelines shown in Table 5, which were adapted from Albrecht et al. (2009).

10	The meaning of the French sentence is fully conveyed in the English translation
8	Most of the meaning of the French sentence is conveyed in the English translation
6	The English translation misunderstands the French sentence in a major way, or has many small mistakes
4	Very little information from the French sentence is conveyed in the English translation
2	The English translation makes no sense at all

Table 5: Adequacy evaluation guidelines for bilingual human judges, adapted from Albrecht et al. (2009).

	Evaluation Category				
	2	4	6	8	10
# sentences	0	0	1	9	231
% sentences	0.0%	0.0%	0.4%	3.7%	95.9%

Table 6: Number and percentage of 241 evaluated sentences judged to be in each category by a bilingual judge. Category labels are defined in Table 5.

The resulting adequacy scores for all 241 post-edited segments are shown in Table 6. We observe that a very high percentage of post-edited segments (95.9% of segments) are rated by the bilingual judge to be completely correct translations of the original French. The remainder are either judged to be mostly correct (3.7% of segments) or partially correct (0.4% of segments). Of the 241 segments, the monolingual post-editor was confident in the post-editing of 209 segments. Those 209 segments were not shown to a bilingual post-editor; we observe that 96.5% of those 209 sentences, which were post-edited by only a monolingual post-editor, are judged to be completely correct by the bilingual judge.

Despite shortcomings (Callison-Burch et al., 2006), BLEU remains a widely used metric for MT evaluation. A somewhat conservative estimate on the quality of the post-edited translations can be measured using BLEU by treating the post-edited translations as a reference translation, and then treating as non-matches (for the purposes of calculating BLEU) all post-edited sentences whose bilingual adequacy score is less than 10; these results are shown in Table 7.

In addition, we cross-tabulate monolingual post-editor confidence (shown in Table 4) with bilingual adequacy judgments (shown in Table 6) to substantiate post-editor confidence with actual translation adequacy for each sentence. The results are shown in Table 8. These results

	BLEU	BLEU-cased
Post-edited (deleting non-perfect translations)	93.3	93.3

Table 7: Translation quality as measured by BLEU (Papineni et al., 2002) of the post-edited machine translation output, treating as non-matches (for the purposes of calculating BLEU) all post-edited sentences whose bilingual adequacy score is less than 10.

		Evaluation Category				
		2	4	6	8	10
Post-Editor Confidence	Completely unsure	0	0	0	0	8
	Partially Unsure	0	0	0	2	11
	Verify	0	0	0	1	10
	Confident	0	0	1	6	202

Table 8: For each of the 241 evaluated sentences, the adequacy category assigned by a bilingual judge, along with confidence assigned by the post-editor. Adequacy category labels are defined in Table 5. Confidence labels are defined in Table 3.

indicate that the high level of post-editor confidence is for the most part justified. Of the 209 segments where the post-editor was confident, only 7 were judged to be less than completely adequate translations.

For reference, these 7 segments are reproduced in their entirety in Appendix A. Four of the segments marked as less than completely adequate contain minor errors in typography or lexical choice. The post-edited translation of Segment 107 substitutes the more technical English term *the data* instead of the more literal *the work* or *the text* for the French term *l'oeuvre*. In segment 171, the English translates the French word *lecteurs* as *readers*, which is a valid translation for that French term, but is an incorrect lexical choice in context. Segment 196 incorrectly uses the literal translation *in the state* instead of a more appropriate idiomatic translation, such as *as is*, for the French phrase *en l'état*. Segment 215 consists entirely of a URL; the post-edited translation is rated 8 instead of 10, presumably because the English "translation" does not faithfully reproduce a spurious space character that appears in the French segment.

The remaining three segments contain more serious problems. The English translation of segment 8 elides a clause present in the French, resulting in an English translation that is perfectly fluent but semantically different from the original French. Segments 183 and 198 each contain phrases that are ill-formed in English, and also do not properly convey the semantic content of the respective French source segments.

4.2 Examining Machine Translation Results

Ideally, it would be desirable to evaluate the raw (un-edited) machine translation using the same 10-point adequacy metric used to evaluate the post-edited translations. Due to time constraints, we were unable to collect bilingual adequacy judgements on the raw (un-edited) machine translation output. We intend to pursue this in future work; in order to enable any interested researchers to perform such an analysis, we are making available for download both the post-edited results and the machine translation output of all three systems as supplementary materials to accompany this paper.

In the absence of a manual evaluation, we consider various automatic metrics in an attempt to provide at least some insight into the machine translation quality. Recall that over 95% of post-edited translations in our task were judged to be completely correct. Given this very high adequacy rate, we propose that it is not unreasonable to treat this post-edited data as reference

		Post-edited translations as reference			
		BLEU	BLEU-cased	PER	WER
System	Google Translate	52.6	51.8	17.70	35.23
	Systran	37.2	36.6	30.29	49.21
	Moses	14.0	11.8	67.34	87.09

Table 9: Similarity of the post-edited translations with the raw (un-edited) machine translation output from each MT system, as measured by case-insensitive and case-sensitive BLEU (Papineni et al., 2002), position-independent word error rate (PER), and word error rate (WER).

		Google translations as reference			
		BLEU	BLEU-cased	PER	WER
System	Google Translate	100.0	100.0	0.0	0.0
	Systran	37.2	36.4	30.2	45.0
	Moses	17.6	15.1	65.6	80.2

Table 10: Similarity of the raw (un-edited) output of Google Translate with the raw (un-edited) machine translation output from the other two MT systems, as measured by case-insensitive and case-sensitive BLEU (Papineni et al., 2002), position-independent word error rate (PER), and word error rate (WER).

		Systran translations as reference			
		BLEU	BLEU-cased	PER	WER
System	Google Translate	37.3	36.6	27.9	41.5
	Systran	100.0	100.0	0.0	0.0
	Moses	21.0	17.9	56.0	72.6

Table 11: Similarity of the raw (un-edited) output of Systran with the raw (un-edited) machine translation output from the other two MT systems, as measured by case-insensitive and case-sensitive BLEU (Papineni et al., 2002), position-independent word error rate (PER), and word error rate (WER).

		Moses translations as reference			
		BLEU	BLEU-cased	PER	WER
System	Google Translate	17.4	14.9	51.2	62.6
	Systran	21.0	17.9	47.4	61.5
	Moses	100.0	100.0	0.0	0.0

Table 12: Similarity of the raw (un-edited) output of Moses with the raw (un-edited) machine translation output from the other two MT systems, as measured by case-insensitive and case-sensitive BLEU (Papineni et al., 2002), position-independent word error rate (PER), and word error rate (WER).

translations in order to examine the quality of the machine translation results used in this experiment. Treating the post-edited translations as a reference translation, we calculate BLEU, word error rate (WER), and position-independent word error rate (PER) on the output from the machine translation three systems.[2]

The results are shown in Table 9. Of the three MT systems, we observe the best scores for the raw MT output from Google Translate. Recall that the monolingual post-editor, when post-editing a segment, had the freedom to use the results of any of the three MT systems as the starting point for post-editing. The good automated metrics scores for Google Translate suggest that the post-editor drew most heavily from the Google Translate results when post-editing.

To get an indication of the relative similarity of the respective segments of the three MT systems, we also calculate BLEU, PER, and WER, treating (in turn) each MT system output as the reference for the purposes of automatic metric calculations. These results are shown in Tables 10, 11, and 12. We observe from these results that the output of Google Translate and Systran are somewhat similar, and that each of those systems differ substantially from the output of Moses.

5 Conclusion

The need for translation in today's highly connected and highly multilingual world far outstrips the supply of qualified human translators. In some cases of assimilation, where a user wants to extract information from a web page or other resource that is in a foreign language, imperfect machine translation can partially or completely satisfy the user's needs. In other more demanding cases of assimilation, as well as in most cases of dissemination, there is a need for a higher quality of translation than most machine translation systems provide.

Monolingual post-editing represents a middle ground between professional translation and raw use of machine translation. Previous work has indicated that monolingual post-editing can result in higher quality results than raw machine translation. In this work we have shown that when the monolingual post-editor is a domain expert in the material being translated, the monolingual post-editor can produce completely correct translations over 95% of the time. This work suggests that a monolingual post-editor can serve to effectively triage the translation process by forwarding on to bilingual post-editors only those segments which are too difficult for the monolingual post-editor to handle.

This work represents an initial examination into monolingual post-editing as a potential triage mechanism for translation. We plan a more thorough examination of this line of research. In future work, we plan to perform manual adequacy evaluations of the raw machine translation output in addition to the post-edited translations, in order to directly measure the adequacy improvements of monolingual post-editing. This work also is limited in scope by only making use of a single monolingual post-editor and a single document; future work should be broader in both of these dimensions, making use of multiple monolingual post-editors (both domain experts and non-experts) and multiple documents to be translated.

Acknowledgements

We wish to thank Katherine Young for her work post-editing, and Jeremy Gwinnup for his work training MT systems. Thanks also to Margaret Stanney and Patricia Phillips-Batoma for their help evaluating translations. Finally, substantial thanks to the anonymous reviewers. Your critiques and comments were extremely helpful, and have made this a better paper.

[2] We calculate BLEU using the `multi-bleu.pl` script from the Moses project, and calculate WER and PER using the `apertium-eval-translator.pl` script from the Apertium project (Forcada et al., 2011). Other metrics more directly tailored for post-editing scenarios, such as Joint Fuzzy Score (Zhechev, 2012) may also be useful to consider in future work.

Appendix

A Segments

- Segment 8 of 241 - Adequacy score 8

  French Les techniques actuelles de traduction automatique (TA) permettent de produire des traductions dont la qualité ne cesse de croitre.

  English Current machine translation (MT) techniques continue to improve.

- Segment 107 of 241 - Adequacy score 8

  French L'oeuvre, composée de 545 segments et 10731 mots est divisée en trois blocs identiques.

  English The data, made up of 545 segments and 10731 words was divided into three equal blocks.

- Segment 171 of 241 - Adequacy score 8

  French Après trois questions permettant de mieux cerner le profil du lecteur, une première partie (5 questions) interroge les lecteurs sur la lisibilité et la qualité du texte littéraire traduit.

  English After three questions to better understand the profile of the player, the first portion (5 questions) asks readers about readability and quality of the translated literary text.

- Segment 183 of 241 - Adequacy score 8

  French ce résultat mitigé indique peut-être un désintérêt de certains lecteurs pour les aspects les plus techniques de l'oeuvre.

  English this mixed result may indicate a lack of interest by some readers to the most technical of the work aspects.

- Segment 196 of 241 - Adequacy score 8

  French Le manque de place ne nous permet pas de commenter ces remarques mais nous pensons qu'elles sont assez explicites pour être délivrées en l'état.

  English Lack of space does not allow us to comment on these remarks but we think that they are sufficiently clear to be delivered in the state.

- Segment 198 of 241 - Adequacy score 6

  French Le texte auquel vous êtes parvenu restitue une image fidéle du contenu de l'article de Powers.

  English The text you have successfully reproduces faithfully the content of the article by Powers.

- Segment 215 of 241 - Adequacy score 8

  French `10. https ://fluidsurveys.com/surveys/manuela-cristina/un-livre-sur-moi-qualite-de-la-traduction/?TEST_DATA=`

  English `10. https://fluidsurveys.com/surveys/manuela-cristina/un-livre-sur-moi-qualite-de-la-traduction/?TEST_DATA=`

References

Albrecht, J. S., Hwa, R., and Marai, G. E. (2009). Correcting automatic translations through collaborations between MT and monolingual target-language users. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 60–68, Athens, Greece.

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.

Besacier, L. (2014). Traduction automatisée d'une oeuvre littéraire: une étude pilote. In *Actes de 21ème Traitement Automatique des Langues Naturelles (TALN '14)*, pages 389–394.

Callison-Burch, C. (2005). Linear B system description for the 2005 NIST MT evaluation exercise. In *Proceedings of the NIST 2005 Machine Translation Evaluation Workshop*.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics (EACL 06)*, page 249256.

Denkowski, M. and Lavie, A. (2012). TransCenter: Web-based translation research suite. In *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice Demo Session*.

Forcada, M., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Google (2014). Google Translate. `http://translate.google.com`.

Green, S., Heer, J., and Manning, C. D. (2013). The efficacy of human post-editing for language translation. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, pages 439–448, Paris, France.

Kazi, M., Coury, M., Salesky, E., Ray, J., Shen, W., Gleason, T., Anderson, T., Erdmann, G., Schwartz, L., Ore, B., Slyh, R., Gwinnup, J., Young, K., and Hutt, M. (2013). The MIT-LL/AFRL IWSLT-2013 MT system. In *The 10th International Workshop on Spoken Language Translation (IWSLT'13)*, pages 136–143, Heidelberg, Germany.

Koehn, P. (2009a). A process study of computer aided translation. *Machine Translation*, 23(4):241–263.

Koehn, P. (2009b). A web-based interactive computer aided translation tool. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 17–20, Suntec, Singapore.

Koehn, P. (2010). Enabling monolingual translators: Post-editing vs. options. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 537–545, Los Angeles, California.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., 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 (ACL '07) Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Langlais, P., Foster, G., and Lapalme, G. (2000). TransType: A computer-aided translation typing system. In *Proceedings of the ANLP/NAACL 2000 Workshop on Embedded Machine Translation Systems*, pages 46–51, Seattle, Washington.

Mitchell, L., Roturier, J., and O'Brien, S. (2013). Community-based post-editing of machine translation content: monolingual vs. bilingual. In *Proceedings of the 2nd Workshop on Post-editing Technology and Practice (WPTP-2)*, pages 35–43, Nice, France. EAMT.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*, pages 311–318, Philadelphia, Pennsylvania.

Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Schwartz, L., Anderson, T., Gwinnup, J., and Young, K. M. (2014). Machine translation and monolingual postediting: The AFRL WMT-14 system. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 186–194, Baltimore, Maryland. Association for Computational Linguistics.

Systran (2010). Systran server 7.4.2. `http://www.systransoft.com`.

Zhechev, V. (2012). Machine Translation Infrastructure and Post-editing Performance at Autodesk. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, pages 87–96, San Diego, USA. Association for Machine Translation in the Americas (AMTA).

Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories

Carlos S. C. Teixeira carlos.teixeira@urv.cat
Translation Studies Research Unit, KU Leuven, Belgium
Intercultural Studies Group, Universitat Rovira i Virgili, Spain
Avda. Catalunya 35, Tarragona, 43002, Spain

Abstract

This paper investigates the behaviour of ten professional translators when performing translation tasks with and without translation suggestions, and with and without translation metadata. The measured performances are then compared with the translators' perceptions of their performances. The variables that are taken into consideration are time, edits and errors. Keystroke logging and screen recording are used to measure time and edits, an error score system is used to identify errors and post-performance interviews are used to assess participants' perceptions. The study looks at the correlations between the translators' perceptions and their actual performances, and tries to understand the reasons behind any discrepancies. Translators are found to prefer an environment with translation suggestions and translation metadata to an environment without metadata. This preference, however, does not always correlate with an improved performance. Task familiarity seems to be the most prominent factor responsible for the positive perceptions, rather than any intrinsic characteristics in the tasks. A certain prejudice against MT is also present in some of the comments.

1. Introduction

Translating as editing of translation memory (TM) matches, on one hand, or as post-editing of machine translation (MT) suggestions, on the other, had traditionally been studied as two separate tasks. However, in recent years research interests have moved to include the language industry's trend of combining translation suggestions from machine translation and translation memories in the same text.

As one would expect, empirical studies with a focus on translation memories (Colominas, 2008; Dragsted, 2004; Garcia, 2007; Moorkens, 2012; Webb, 1998) have reported on the use of typical translation memory systems. These are tools that offer one or more translation suggestions as the user activates a segment and that always display metadata about those suggestions, i.e. they indicate where the suggested translations come from, how similar to the reference source segment the current source segment is (fuzzy match level) and where the textual differences lie. In contrast, studies on pure machine translation post-editing (Allen, 2003; Almeida, 2013; Garcia, 2011; Guerra Martínez, 2003; Krings, 2001; Plitt & Masselot, 2010) have often resorted to editing environments that offer pre-translated text with no associated metadata, as this is the typical setup for such tools. Yet the scenario for post-editing is starting to change with the development of post-editing environments that can display confidence estimates for machine translation suggestions, such as PET (Aziz, Sousa, & Specia, 2012) and CASMACAT (2014). Those estimates are believed to represent useful metadata for repairing MT suggestions.

Some studies have compared unaided translation with TM-assisted translation or with MT-assisted translation. A recent example of the latter is Green, Heer, and Manning (2013), in

which the authors take into account the translators' perceptions by means of questionnaires, like we do in the current paper. However, only a few studies have analysed scenarios in which machine translation and translation memories are combined in the same workflow. These studies either use existing TM systems (O'Brien, 2006; Skadiņš, Puriņš, Skadiņa, & Vasiļjevs, 2011; Yamada, 2011) or they resort to a purpose-built post-editing environment (Guerberof, 2009; He, Ma, Roturier, Way, & van Genabith, 2010), as there seem to be no established tools for post-editing. One question that arises from this dichotomy is how to compare the performance of TM suggestions against MT suggestions in an environment that has not been conceived with their integration in mind. On the one hand, in a post-editing tool TM matches are analysed without the associated metadata, which are an important feature of translation memory systems (Anastasiou & Morado Vázquez, 2010; Karamanis, Luz, & Doherty, 2011; Morado Vázquez, 2012; Teixeira, 2014) but are not present in post-editing tools. Metadata allow translators not only to make choices among different types of suggestions, but also to decide how to approach a suggestion when repairing it. On the other hand, in a traditional TM system, MT suggestions have to be manually inserted in the active segment and are presented surrounded by much more information than is typical in a post-editing tool, maybe decreasing the translation speed for this suggestion type and increasing the post-editor's cognitive load. Therefore, comparing the performances of TM vs. MT suggestions is not an easy task, as the general tendency is to assess one of the suggestion types in an environment for which it was not originally intended to be used.

The current paper seeks to consider this issue while investigating certain aspects of TM/MT integration. It focuses on metadata and pre-translation as control variables, and analyses how they affect translators' performances and perceptions. The study reported on here uses a traditional TM system, but the system is set up using different configurations, in an attempt to "favour" one suggestion type at a time: one task reproduces an environment that is more typical of TM systems – interactive translation (Wallis, 2006) with metadata –, while the other task is more typical of MT post-editing tools – pre-translation with no metadata.

The participants' performances are measured in terms of time, edits and errors. Time and edits are measured using keystroke logging tools, while the errors are assessed by two professional reviewers using an error-score system. This measured data is triangulated with perception data obtained from interviews done with each translator immediately after the translation tasks. The goal of this triangulation is to analyse how the presence or absence of priming elements such as suggested translations and metadata affect translators and to determine whether those factors could be the main determinants for any differences found in performance.

2. Experiment description

An experiment was run with ten professional translators working from English into Spanish, who performed three different tasks within the same tool. One task presented no translation suggestions (translation from Scratch); another task presented translation suggestions from both TM and MT, and metadata about the suggestions (Visual task); and another task presented pre-translated text also from TM and MT but no metadata about the suggestions (Blind task).

2.1. Participants

The ten translators who took part in the experiment were native speakers of Spanish, with some of them being bilingual Spanish/Catalan speakers. There were five men and five women, with ages ranging from 24 to 51. They had been working for 1.5 to 18 years as full-time translators for a small translation company in Barcelona, where they had been translating IBM material

and using IBM TranslationManager[1], the translation memory system used in the experiment. They all had experience post-editing machine translated texts for IBM and/or other customers for 0.5 to 3 years. As a compensation for performing the tasks in the experiment, they were paid their regular hourly rates. Table 1 shows the demographics of the experiment participants.

Participant	Gender	Age	Years working as a translator	Years working with IBM TM/2	Years working with MT post-editing
P01	F	30	7	6	0.5
P02	M	37	14	13	0.5
P03	F	32	3.5	3	0.5
P04	M	26	2.5	2	2.0
P05	F	26	3	3	0.3
P06	M	29	2.5	2	0.5
P07	F	24	1.5	1.5	1.0
P08	M	51	18	18	0.8
P09	F	43	10	10	3.0
P10	M	47	15	14	0.5

Table 1: Demographic data about participant translators

2.2. Translation tasks

For the sake of ecological validity, the experiment was conducted with translators working with their computers of habitual use in their regular office space, and the project was configured in a way as similar as possible to their normal IBM assignments. Each translator was asked to perform the following three tasks:

a) Translation from Scratch: To translate a short text (118 words, 5 segments) from English into Spanish in IBM TranslationManager, without any help from translation memories or machine translation.

b) Translation in a Visual setting: To translate a longer text (505-542 words, 28 segments) from English into Spanish in IBM TranslationManager, with one translation suggestion per segment and metadata about the translation suggestions.

c) Translation in a Blind setting: To translate a longer text (505-542 words, 28 segments) from English into Spanish in IBM TranslationManager, with pre-translated segments but no metadata about the translation suggestions.

Task *a* (Scratch) was always the initial task, while Tasks *b* (Visual) and *c* (Blind) were performed in different orders depending on the participants, in order to have an even distribution of task orders. Task *a* was always first because it served rather as a warm-up activity and was not the focus of the study. Two different texts were used for Tasks *b* and *c* and distributed evenly between the two tasks. Table 2 shows the distribution of task and text orders among the participants.

The source texts used for the three translation tasks were excerpts from the *Troubleshooting Guide* for the IBM Tivoli Monitoring software. In the task where translators had to type from scratch, a text with 118 words and 5 segments was used, and no translation suggestions were provided. Translators were instructed to open a previously configured folder (project) in IBM TranslationManager and to translate the only file it contained.

[1] Also known as TM/2

	1st Task		2nd Task		3rd Task	
Participant	Configuration	Text	Configuration	Text	Configuration	Text
P01	Scratch	0	Blind	1	Visual	2
P02	Scratch	0	Blind	2	Visual	1
P03	Scratch	0	Visual	2	Blind	1
P04	Scratch	0	Visual	2	Blind	1
P05	Scratch	0	Blind	2	Visual	1
P06	Scratch	0	Blind	2	Visual	1
P07	Scratch	0	Blind	1	Visual	2
P08	Scratch	0	Blind	1	Visual	2
P09	Scratch	0	Visual	1	Blind	2
P10	Scratch	0	Visual	1	Blind	2

Table 2: Distribution of task and text orders among the participants

For the Visual and Blind tasks, each of the 28 segments in the texts was randomly assigned one of four possible types of translation suggestions – exact matches, fuzzy matches in the 70-84% range, fuzzy matches in the 85-99% range and machine translation feeds – resulting in seven translation suggestions of each type per text. An authentic IBM translation memory was used as a reference for producing the exact and fuzzy matches, without any special tricks being inserted intentionally. The machine translation feeds came from a commercial Moses (Koehn et al., 2007) statistical engine that had been trained with product-specific terminology and was used in production for regular IBM projects in the company.

In the Visual task, one translation suggestion was provided for each segment, and the translators had to actively insert it in the editing area and edit it if they considered it to be a usable suggestion, or they could type their translation either from scratch or on top of the source text. The most common way for the translators to insert translation suggestions was by using a keyboard shortcut, although in some cases they preferred to copy and paste either the whole or parts of the suggestions. In this task, translation suggestions were provided with metadata, which in IBM TranslationManager are indicated by means of a letter placed to the left of the suggestion: blank for exact matches, "f" for fuzzy matches and "m" for machine translation feeds. Additionally, in the case of fuzzy matches, the tool highlights the text portions that differ between the source text in the active segment and the source segment in the translation memory.

In the Blind task, there was also one translation suggestion per segment, but the suggestion had been previously inserted in the segment, so the file displayed as pre-translated text to be edited, instead of source text to be replaced with a translation suggestion. The application panes where the translation suggestions are usually displayed were empty, so no translation metadata were displayed.

2.3. Interviews

The interviews were conducted immediately after the translation tasks, both as semi-structured dialogues and as retrospection with replay (see Hansen, 2008). The base questions asked during the dialogues were:

1) Do you think you translated *faster* in any of the environments? If so, in which one?
2) Do you think the quality of your final translation was *better* in any of them? If so, in which one?
3) In which environment did you feel more *comfortable* working?

During the retrospection, the translators watched selected passages from their performance recordings and commented on certain aspects of the translation tasks based on prompts

from the researcher. For two participants it was not possible to carry out the retrospection, because of technical reasons (P05) and because one participant refused to do it (P08).

3. Data collection and processing

The translation processes were recorded with BB FlashBack and Inputlog (Leijten & van Waes, 2013). This made it possible to measure the total time spent and the total number of characters typed by each translator in each task. All translations were then assessed for quality by two reviewers, who had been revising this type of material for 12 and 19 years in the company. The reviewers revised the translations as Word documents by highlighting their corrections with the *Track Changes* feature. The severity of errors had been previously identified through a series of interviews with project managers in the company, based on their common practice for this type of translation project. Errors related to misinterpretation of the original, missing or added information, tag corruption and misspelt brand names scored two points. Errors such as inconsistencies, misspellings, wrong grammar and punctuation scored one point. Other text issues such as those related to style and fluency were not taken into account. The researcher acted as a third reviewer, making small adjustments to the scores when the two reviewers had too different opinions and marking any obvious errors that had not been detected by the reviewers.

As for the qualitative data, the interviews were recorded then transcribed and coded. In order to better visualise the results, tables were created for each subject, where the verbal data was organised according to the three tasks (Scratch, Visual, Blind) and the three main variables: time (verbalised as 'speed'), effort (verbalised as 'comfortable') and quality.

A third and last data analysis step was necessary to make the qualitative and quantitative data comparable. The approach used here was to rank each variable in each of the tasks for each subject, both as measured and as perceived, and then to compare the rankings. The next section explains this method and presents the results.

4. Results and analysis

4.1. Quantitative data

Table 3 shows the measured results for all ten subjects. *Time* is indicated as seconds per 100 source words. *Edits* is a percent ratio between the total number of relevant key presses and the total number of characters in the final target text, including spaces. *Errors* is the total number of weighted errors (as explained in the previous section) per 100 source words.

	TIME			EDITS			ERRORS		
Participant	Scratch	Visual	Blind	Scratch	Visual	Blind	Scratch	Visual	Blind
P01	257	191	200	102	14.0	11.9	3.8	1.0	1.0
P02	235	167	229	97	22.8	15.7	1.3	1.9	1.4
P03	324	215	193	103	50.0	13.3	5.5	4.3	4.5
P04	566	223	266	103	16.8	15.4	3.8	3.1	3.6
P05	259	121	157	106	12.4	11.5	5.1	4.3	4.2
P06	296	143	162	102	12.0	12.1	4.2	4.8	5.0
P07	613	232	334	109	13.0	14.5	3.8	3.3	3.1
P08	777	497	343	132	29.8	18.6	3.0	1.2	1.9
P09	344	139	139	153	22.6	6.37	8.9	5.4	5.0
P10	240	139	120	108	16.1	9.32	3.0	4.3	5.2

Table 3: Measured times (seconds/100 words), edits (%) and errors (weighted errors/100 words) per participant in the three translation tasks

In Table 4, the values shown in Table 3 are converted into score levels. Thus, for each particular subject and for each variable in Table 3, the task with the lowest number is assigned level 1 in Table 4, the task with the highest number is assigned level 3 and the intermediary task is assigned level 2. When the difference between two tasks is not relevant, considering a deviation of ± 5 percent, the same level is assigned to more than one task, giving preference to the extreme levels 1 and 3. The reason for preferring the extremes is that it corresponds better to human perception and to the types of answers available from the interviews (e.g. the fastest task vs. the slowest task).

	TIME			EDITS			ERRORS		
Participant	Scratch	Visual	Blind	Scratch	Visual	Blind	Scratch	Visual	Blind
P01	3	1	1	3	2	1	3	1	1
P02	3	1	3	3	2	1	1	3	1
P03	3	2	1	3	2	1	3	1	1
P04	3	1	2	3	2	1	3	1	3
P05	3	1	2	3	1	1	3	1	1
P06	3	1	2	3	1	1	1	3	3
P07	3	1	2	3	1	2	3	1	1
P08	3	2	1	3	2	1	3	1	2
P09	3	1	1	3	2	1	3	1	1
P10	3	2	1	3	2	1	1	2	3

Table 4: Measured times, edits and errors as a score level in the three translation tasks

Table 4 indicates that all translators spent the most time and made the most edits (represented by the number 3) when translating from Scratch. The same cannot be said about the errors, since three of the translators made the fewest errors when translating from Scratch. The table also shows that most translators performed the fewest edits in the Blind task, except for one translator, who typed less in the Visual task. More will be said about the results in this table when comparing them with the translators' perceptions.

4.2. Qualitative data

Table 5 shows how the translators perceived their performance after the translation tasks, as a result of coding the interview data.

	TIME			"EFFORT"			ERRORS		
Participant	Scratch	Visual	Blind	Scratch	Visual	Blind	Scratch	Visual	Blind
P01	3	3	1		1		1	1	1
P02	3	1	1		1	2		1	1
P03		1		3	1	1	3	1	1
P04		1			1			1	3
P05		1			1			1	
P06	2	1	3	2	1	3	2	1	3
P07		1		2	1	3	2	1	3
P08	3	2	1		1	1		1	2
P09		1		2	1	3			3
P10		1	2		1			1	1

Table 5: Perceived time, effort and errors as a score level in the three translation tasks

The blank cells in the table represent data for which no clear answer was given in the interview. As a general observation, the table shows that all participants thought they made

fewer errors and invested less effort in the Visual task than in any of the two other translation tasks, and that most of them considered they spent the least time on the Visual task. In the following sections, we will compare the measured and perceived data in detail for each of the dependent variables.

4.3. Comparison between quantitative and qualitative data

The time measured per 100 words was consistently higher when translating from scratch for all ten participants. This is in accordance with their perception, except for one translator, who thought he spent less time translating from scratch than he did in the Blind task. For the seven translators who thought they were faster in the Visual task than in the Blind task (P03, P04, P05, P06, P07, P09, P10), all but two (P03 and P10) were indeed faster. For the two translators who thought they were faster in the Blind task than in the Visual task (P01, P08), their perception corresponded to their measured times. The only participant who thought he was as fast in the Visual as in the Blind task (P02) was actually much faster in the Visual task. P01 thought she spent the least time on the Blind task, whereas she actually spent less time on the Visual task.

Seventy percent of translators made the most errors when translating from scratch, which might indicate their reliance on translation suggestions, after many years of practice working with translation memories. There was no clear advantage between the Visual and the Blind tasks in terms of error rates, although all the translators thought they made the fewest errors in the Visual task, except for one translator, who did not distinguish explicitly between the Visual task and translating from scratch. Their perception corresponded with the reviewers' quality assessment in 70 percent of the cases, whereas two translators actually made the most errors in the Visual task and one translator made more errors in the Visual task than when translating from scratch.

As indicated in Table 4, the Blind task was the condition in which the translators typed the least, except for one translator, who typed less in the Visual task. Two translators typed as much in the Blind task as in the Visual task. A simple comparison of the middle columns in Table 4 and Table 5 reveals no coincidence between the measured edits and the perceived "effort" while performing the task. This could be attributed to any of the factors mentioned in Section 4.4, but in this case, the discrepancies in the results are probably due to a poorly formulated question. The quantitative variable being measured as an indication of effort was the amount of editing, which is a simple measurement of physical effort, while in the interviews the translators were asked about the task in which they felt more "comfortable". It turns out that typing effort and the feeling of "comfort" while performing a task are not directly comparable. This is in accordance with the conclusions of other studies, such as Koponen, Aziz, Ramos, and Specia (2012), who suggest that "keystrokes, while very useful as a way to understand how translators work, may not be an appropriate measure to estimate cognitive effort" (p. 20).

4.4. Additional information from the interviews

A major goal of the interviews was to let participants express their priorities. This was achieved through a relatively free dialogue format, which was responsible for some missing data in Table 5, but also allowed other factors to come into play that had not been included as the main variables in the study.

Translation vs. revision vs. post-editing

The interviews indicate a clear difference in the way translators perceived the two main translation tasks. All participants except one made a clear distinction between "translate", for the Visual task, and "revise" or "proofread" ("revisar", in Spanish) or "post-edit", for the Blind

task.[2] The quantitative data support this perception, as they show many more iterations per segment in the Visual environment, as if the translators were first translating, then self-revising. In the Blind environment, which they considered to be revising or post-editing, they completed the task in a single round. This difference made seven of the translators feel that they had performed a regular revision (on text that had been translated or proofread by another translator) when working in the Blind task (my translations here and throughout):

> P10: I'm very much used to working the first way, to translate. I had never done the other task before actually, to find everything at 100% and to revise it.

> P02: The other one was already done, we just had to revise.

> P01: Post-editing, a revision that had already been done and that I had to revise.

For these participants, the text they were "revising" was in principle better than the text they had in the Visual task:

> P08: We assume that in theory it should be better.

> P04: There was a lot of [translation] memory and it was quite good compared with other folders.

> P07: We could notice some segments had been leveraged from the memory... they were better, I didn't have to change much.

Only one participant felt she was "translating" when performing the Blind task: she actually talked about both tasks in terms of the presence or absence of metadata on the translation suggestions (P05).

The role of translation suggestions

Seven translators acknowledged the usefulness of translation suggestions (as opposed to translating from scratch):

> P02: Because [when you translate from scratch] you have to think more.

> P03: It always helps to have pre-translated stuff or when there is something previous that is useful, because if you translate everything from scratch, you always make mistakes, [it's a little] more difficult. Having something as a basis is always welcome.

> P06: When you have a suggestion from the memory, you insert it and if you change a word, maybe you go faster too, with some memory. [Pause] Translating 500 words with memory suggestions is faster than from scratch…

> P07: Because you have an external aid from previous memories and machine translation [...] you always go faster. [...] it is always better to have some help.

One of those participants (P09), however, pondered that it might be easier to translate from scratch:

[2] In the current state of play, with MT and TM suggestions being presented together, it is not surprising that no clear distinction is made between post-editing and revising.

P09: It is easier to translate from scratch, because I don't have to look at anything. And I don't need to check if what is suggested is correct or not, or if it's in the right order or in the wrong order.

Along the same lines, P01 said:

P01: I don't think it is especially faster having the memory, because when you translate from scratch, one advantage I can see is the vocabulary, but the other is that there is no suggestion to look at, no differences to check for between one sentence and the other. [...] I think I compensate what I use—the help from the memory—with the time I spend checking the passage, checking for differences.

The role of metadata

Even if the translators did not consider the metadata to be the main distinction between the Visual and the Blind tasks in their comments, they demonstrated awareness of how translation metadata could help them:

P01: If I see a fuzzy match, the first thing I'll look at is the Source of Proposal. For me it's easier with a memory, with fuzzy matches, with information on whether it comes from MT or from fuzzy or whatever, because it allows me to look at it in one way or another.

P02: If you see that it's 100%, that it's not machine translation, then, in principle, in an everyday translation, when you go fast, you don't even look at it. You assume it's correct or that you translated it yourself before [...] A fuzzy match, if I see that everything is translated and there is only one word that changes, I change that word, I don't even look at the rest.

P04: Because you can't see below where it comes from... [when there is no metadata]

P05: TM/2 indicates the fuzzy matches... it highlights what is missing, what is extra, what has changed.

P06: You always look at what has changed and you change there. [...] You didn't even need to read the sentence, you just had to change a word that was highlighted and that's it.

P08: When it's pre-translated you don't have... you don't know the quality of the suggestion; in contrast, when you have the memory, you know if it's an MT suggestion or if it comes from a... from another publication. TM/2 indicates if it's an Exact Match or if it's an MT suggestion or if it's a fuzzy match... [...] Sometimes you just look at what has changed. On the other hand, when you have it pre-translated, I don't know where it comes from... I would prefer to know... the environment where you see the suggestion, if it's machine translation, if it's... or if it comes from another publication that has been checked by somebody else. I think it's better to have the information, because it tells you what has changed; so if you know what's changed, you focus more on what's changed. Your natural tendency is to trust more what appears as unchanged.

P09: The second one [Visual] had several fuzzies at 95%, 85%, so it's very easy to detect where the small changes are, and it's very useful. [...] If you look at the suggestion, since it tells you exactly what the changes are, it's easier to detect. [...] For me it's much easier to upload or to edit.

Morado Vázquez (2012) obtained similar feedback from the translators in her study: "In terms of participants' attitude towards the metadata received, most of the participants did not find it distracting, and the majority of them would prefer a translation memory which contained metadata." It is worth noting, however, that one of my translators stated, "the environment that gives you more information is, at the same time, more complex" (P08).

The perception of machine translation

In general, the participants had mixed feelings about machine translation. Although in some cases they criticised it as being poor, they also recognised that some machine-translated segments were "almost perfect" and that MT helped them increase productivity.

Two translators felt the text in the Blind task contained more machine-translated segments than the text in the Visual task, although the translators were told that both texts actually had the same distribution of suggestion types, and only 25% of the suggestions were actually machine translation feeds (see Section 2.2). Therefore, in their comments the translators made statements about the (presumably lower) quality of the translation suggestions based on their assumption that the suggestions came from machine translation:

> P06: In the revision task, since they come from machine, they are always faulty.

> P09: [The Blind task] is mostly machine, so it takes me longer to think about what changes [...]. I do have to keep thinking what the core of the segment is and to change it.

He et al. (2010) and Guerberof (2013, pp. 87–88) also show evidence that translators tend to trust fuzzy matches more than they trust machine translations and that in many cases subjects are not able to tell TM suggestions from MT suggestions.

Task familiarity

Eight out of the 10 participants (P01, P02, P04, P05, P06, P07, P09, P10) reported being more comfortable tackling the Visual task, even when some believed the Blind task could be faster. The other two participants (P03 and P08) were equally comfortable working in the Blind task. P08 found the Blind task "more simple":

> P08: You look at the English, the Spanish and that's it. [...] In the other one, you have to look at the English, the Spanish, and sometimes choose among five suggestions – not the case in this experiment though, where you had only one suggestion.

The main reason given by the translators (mentioned by 7 out of 10) for feeling more comfortable and actually preferring the Visual task was that they were very "used to" or "more familiar with" (in Spanish: "acostumbrado", "familiarizado", "habituado") the Visual task, while the Blind task was new to them. Another reason given by the translators (3 out of 10) for preferring the Visual task was that they felt more confident in this environment. It is unclear in some statements whether this feeling of confidence is only related to task familiarity or also to the metadata or to any other characteristics present in the Visual task.

> P01: I prefer to translate with a memory. [...] For me it's more comfortable, it makes me feel more confident.

> P04: Surely because this is what I've been doing for IBM lately, [I feel] more confident, maybe more familiar with it.

P08: If you know it's a fuzzy match, since you know it has been checked by a human translator, it gives you more confidence.

Different strategies

Since all the participant translators were used to doing revisions in IBM TranslationManager, where the text to be revised comes pre-translated (but with metadata on the provenance of existing translations), their feeling of unfamiliarity or lack of confidence with the Blind task can probably be explained by the absence of metadata in this task. This suspicion is reinforced by several statements in which translators explain that they use different strategies for exact matches, fuzzy matches and machine translation:

P01: If I see it's an "m" [machine translation], I read the sentences from A to Z, or I go and check for some things or I look for some things or for other things. If I see a fuzzy match, the first thing I'll look at is the Source of Proposal. For me it's easier with a memory, with fuzzy matches, with information on whether it comes from MT or from fuzzy or whatever, because it allows me to look at it in one way or the other. If I see a fuzzy match, I look at the Source of Proposal; if I see an MT, that is, if I see an "m", and it gives me the impression that the sentence is more or less correct, then I insert it and, depending on the case, I fix it, because sometimes the sentence is almost entirely perfect.

P02: If you know it's... you look at it differently. If you see that it's 100%, that it's not machine translation, then, in principle, in an everyday translation, when you go fast, you don't even look at it.

P08: [...]if you know it's MT, you look at it with more... respect. Conversely, if you know it's a fuzzy match, since you know it has been checked by a human translator, it gives you more confidence. Sometimes you just look at what has changed.

These testimonials are in accordance with feedback provided by participants in other studies (O'Brien, 2006, p. 198), as different types of translation tasks seem to activate different translation strategies and to require different allocation of cognitive resources (Carl, Kay, & Jensen, 2010; Dragsted, 2012; House, 2000; Hvelplund, 2011; Jääskeläinen, 1993; Lörscher, 1991). The fact of knowing which type of suggestion is being dealt with when processing a segment could reduce cognitive load and account for the reported feeling of comfort.

5. Discussion

Although the quantitative results between the three environments do not show a clear advantage when translating a specific task, participants preferred to work on the more traditional Visual task, with translation suggestions and metadata. This might be explained by a feeling of increased performance in some cases, as they tended to over-rate the Visual task, but also by task familiarity and the increased level of confidence resulting therefrom.

The metadata factor (present in the Visual task, absent in the Blind task) did not correlate with a consistent increase in performance according to the measured data. A more in-depth analysis of the experiment results has shown that this factor does have a positive effect on performance indicators for certain types of translation suggestions, namely high fuzzy matches and exact matches. The results presented here indicate that metadata are also a relevant factor to increase confidence and reduce cognitive load, by giving translators a hint on how to initially approach a suggestion, as they reportedly use different strategies for different kinds of suggestions.

In the current experiment, only one translation suggestion was presented for each segment, so the study only allowed us to analyse how metadata can help translators use the one suggestion provided. Since translating with CAT tools usually involves a dual process of selection + repairing of suggestions, it would be interesting to complement the current study with a follow-up experiment including multiple suggestions, to investigate how metadata can also help translators choose among different proposals. Likewise, the experiment could be extended by isolating the "pre-translation" and "metadata" factors, as in the current study both those variables were playing a role: one task had pre-translation and no metadata and the other one had "regular translation" and metadata.

The pre-translation factor has also proved to affect translators psychologically in the way they approached the text and the trust they attributed to the proposals – having being previously translated by an (assumedly reliable) human translator.

In the interviews, the question "In which environment did you feel more comfortable?" assumed that "comfortable" (Spanish "cómodo") might inversely correlate with typing effort. This proved to be a very naive assumption, as comfort seems to correlate more with long-time experiential factors than with momentary task characteristics. If a similar experiment is reproduced, the question to be asked should be simply "In which environment do you think you typed more?". Alternatively, a different measurement for cognitive effort should be used.

Still regarding the interviews, a better strategy should be found to elicit answers for the variables in all tasks, in order to have all cells completed in Table 5, while still making sure the answers are not influenced by the researcher's prompts. The interview data in this study are admittedly incomplete, but they have still provided enough information to draw relevant conclusions about the translators' perceptions.

6. Conclusion

The goal of this paper was two-fold: first, to propose a translation environment where suggestions coming from a translation memory and from machine translation could be compared on a fair basis; second, to compare the measured performances and perceived performances of professional translators when exposed to different translation conditions.

The first goal was pursued by setting up two tasks in the same tool, one that emulated a typical TM-assisted workflow and another one that was more typical of post-editing environments. Some problems were found and the task setups should still be improved in future studies to bring both tasks closer to real scenarios.

The second goal was pursued by ranking the measured performances, ranking the perceived performances and comparing both rankings. Not all expected answers could be elicited during the interviews, but the missing data did not prevent us from making conclusive observations. The main conclusion is that translators' perceptions about their performances do not always correlate with their actual performances. The interviews also provided additional information on topics such as task familiarity and translation strategies, indicated that translators tend to associate pre-translated text with revision and post-editing, and gave hints on the translators' opinions about machine translation.

The study found that the measured performances were positively affected by the presence of translation suggestions, but not so much by the presence of translation metadata. However, the interviews indicate that translators preferred the task with translation metadata, even when it did not correlate with an improved performance. Most of the participants felt more comfortable handling this task and had the impression it allowed them to work faster and to make fewer errors. The main reason identified for the positive perception of the Visual task was task familiarity.

A general correlation between being familiar with a task and preferring to do that task is not a particularly surprising result. Indeed, it seems to follow a general trend related to the

adoption of new technologies, as previously reported by studies such as Dillon and Fraser (2006). However, it might suggest that practice is a major factor to improve job satisfaction, even if it does not always imply increased performance.

Acknowledgments

I would like to thank Anthony Pym and three anonymous reviewers for their comments on earlier versions of this manuscript. I would also like to acknowledge the funding to my doctoral research, provided through the European Commission's TIME Marie Curie fellowship (FP7-PEOPLE-2010-ITN-263954).

References

Allen, J. H. (2003). Post-editing. In H. L. Somers (Ed.), *Computers and translation. A translator's guide* (pp. 297–317). Amsterdam, Philadelphia: John Benjamins Pub. Co.

Almeida, G. de. (2013). *Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two Romance languages* (Doctoral thesis). Dublin City University, Dublin. Retrieved from http://doras.dcu.ie/17732/

Anastasiou, D., & Morado Vázquez, L. (2010). Localisation Standards and Metadata. In S. Sánchez-Alonso & I. N. Athanasiadis (Eds.), *Communications in Computer and Information Science. Metadata and Semantic Research* (pp. 255–274). Berlin, Heidelberg: Springer Berlin Heidelberg.

Aziz, W., Sousa, S., & Specia, L. (2012). PET: A Tool for Post-editing and Assessing Machine Translation. In : *LREC, Eighth International Conference on Language Resources and Evaluation* (pp. 3982–3987). Instanbul, Turkey. Retrieved from http://www.mt-archive.info/LREC-2012-Aziz.pdf

Carl, M., Kay, M., & Jensen, K. T. (2010). *Long Distance Revisions in Drafting and Post-editing: Paper presented at CICLing-2010, Iaşi, Romania*. Retrieved from http://research.cbs.dk/portal/en/publications/long-distance-revisions-in-drafting-and-postediting%28fd3ffefc-6ea1-4362-9fc4-be80fef79af7%29/export.html

CASMACAT. (2014). *Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation*. Retrieved from http://www.casmacat.eu/

Colominas, C. (2008). Towards chunk-based translation memories. *Babel*, *54*(4), 343–354. doi:10.1075/babel.54.4.03col

Dillon, S., & Fraser, J. (2006). Translators and TM: An investigation of translators' perceptions of translation memory adoption. *Machine Translation*, *20*(2), 67–79. doi:10.1007/s10590-006-9004-8

Dragsted, B. (2004). *Segmentation in Translation and Translation Memory Systems: An empirical investigation of cognitive segmentation and effects of integrating a TM system into the translation process* (Doctoral thesis). Copenhagen Business School, Frederiksberg.

Dragsted, B. (2012). Indicators of difficulty in translation — Correlating product and process data. *Across Languages and Cultures*, *13*(1), 81–98. doi:10.1556/Acr.13.2012.1.5

Garcia, I. (2007). Power shifts in web-based translation memory. *Machine Translation*, *21*(1), 55–68. doi:10.1007/s10590-008-9033-6

Garcia, I. (2011). Translating by post-editing: is it the way forward? *Machine Translation*, *25*(3), 217–237. doi:10.1007/s10590-011-9115-8

Green, S., Heer, J., & Manning, C. D. (2013). The efficacy of human post-editing for language translation. In W. E. Mackay, S. Brewster, & S. Bødker (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 439–448).

Guerberof, A. (2009). Productivity and quality in the post-editing of outputs from translation memories and machine translation. *Localisation Focus - The International Journal of Localisation*, *7*(1), 11–21.

Guerberof, A. (2013). What do professional translators think about post-editing? *The Journal of Specialised Translation*, (19), 75–95. Retrieved from http://www.jostrans.org/issue19/art_guerberof.php

Guerra Martínez, L. (2003). *Human Translation versus Machine Translation and Full Post-Editing of Raw Machine Translation Output* (Master's thesis). Dublin City University, Dublin.

Hansen, G. (2008). The dialogue in translation process research. In *Translation and Cultural Diversity. Selected Proceedings of the XVIII FIT World Congress 2008* . Shanghai, China: Foreign Languages Press. Retrieved from http://www.translationconcepts.org/pdf/Hansen_ArticleMethods.pdf

He, Y., Ma, Y., Roturier, J., Way, A., & van Genabith, J. (2010). Improving the Post-Editing Experience using Translation Recommendation: A User Study. In *AMTA 2010 - 9th Conference of the Association for Machine Translation in the Americas* . Retrieved from http://amta2010.amta-web.org/AMTA/papers/2-27-HeMaEtal.pdf

House, J. (2000). Consciousness and the Strategic Use of Aids in Translation. In S. Tirkkonen-Condit & R. Jääskeläinen (Eds.), *Tapping and mapping the processes of translation and interpreting. Outlooks on empirical research* (pp. 149–162). Amsterdam/Philadelphia: John Benjamins.

Hvelplund, K. T. (2011). *Allocation of cognitive resources in translation. An eye-tracking and key-logging study* (Doctoral thesis). Copenhagen Business School, Frederiksberg.

Jääskeläinen, R. (1993). Investigating translation strategies. In S. Tirkkonen-Condit & J. Laffling (Eds.), *Kielitieteellisiä tutkimuksia, Studies in languages: Vol. 28. Recent trends in empirical translation research* (pp. 99–120). Joensuu: Joensuu University. Retrieved from http://scholar.google.com/scholar?cluster=8725738219265795995&hl=en&as_sdt=0,22

Karamanis, N., Luz, S., & Doherty, G. (2011). Translation practice in the workplace: contextual analysis and implications for machine translation. *Machine Translation*, *25*(1), 35–52. doi:10.1007/s10590-011-9093-x

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., . . . 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* (pp. 177–180). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P07-2045

Koponen, M., Aziz, W., Ramos, L., & Specia, L. (2012). Post-editing time as a measure of cognitive effort. In : *WPTP, AMTA 2012 Workshop on Post-Editing Technology and Practice* (pp. 11–20). San Diego, USA. Retrieved from http://www.mt-archive.info/AMTA-2012-Koponen.pdf

Krings, H. P. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes.* (Koby, G. S., Ed.). Kent, Ohio: Kent State University Pr.

Leijten, M., & van Waes, L. (2013). Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication*, *30*(3), 358–392. doi:10.1177/0741088313491692

Lörscher, W. (1991). *Translation performance, translation process and translation strategies: A psycholinguistic investigation*. Tübingen: Gunter Narr.

Moorkens, J. (2012). *Measuring Consistency in Translation Memories. A Mixed-Methods Case Study* (Doctoral thesis). Dublin City University, Dublin.

Morado Vázquez, L. (2012). *An empirical study on the influence of translation suggestions' provenance metadata* (Doctoral thesis). University of Limerick, Limerick.

O'Brien, S. (2006). Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology*, *14*(3), 185–205. doi:10.1080/09076760708669037

Plitt, M., & Masselot, F. (2010). A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, *93*, 7–16. doi:10.2478/v10108-010-0010-x

Skadiņš, R., Puriņš, M., Skadiņa, I., & Vasiļjevs, A. (2011). Evaluation of SMT in localization to under-resourced inflected language. In M. L. Forcada, H. Depraetere, & V. Vandeghinste (Eds.), *Proceedings of the 15th conference of the European Association for Machine Translation* (pp. 35–40).

Teixeira, C. S. C. (2014). The handling of translation metadata in translation tools. In S. O'Brien, L. Balling, M. Carl, M. Simard, & L. Specia (Eds.), *Post-editing of machine translation. Processes and applications* (pp. 109–125). Newcastle upon Tyne: Cambridge Scholars Publishing.

Wallis, J. (2006). *Interactive Translation vs Pre-translation in the Context of Translation Memory Systems. Investigating the effects of translation method on productivity, quality and translator satisfaction* (Master's thesis). University of Ottawa, Ottawa. Retrieved from http://www.localisation.ie/resources/Awards/Theses/Thesis%20-%20Julian%20Wallis.pdf

Webb, L. E. (1998). *Advantages and Disadvantages of Translation Memory. A Cost-Benefit Analysis* (Master's thesis). Monterey Institute of International Studies, Monterey.

Yamada, M. (2011). *Revising text: An empirical investigation of revision and the effects of integrating a TM and MT system into the translation process* (Doctoral thesis). Rikkyo University.

Perception vs Reality: Measuring Machine Translation Post-Editing Productivity

Federico Gaspari fgaspari@computing.dcu.ie
CNGL, School of Computing, Dublin City University, Dublin 9, Ireland
Antonio Toral atoral@computing.dcu.ie
CNGL, School of Computing, Dublin City University, Dublin 9, Ireland
Sudip Kumar Naskar * sudip.naskar@cse.jdvu.ac.in
Jadavpur University, Kolkata, India
Declan Groves * degroves@microsoft.com
Microsoft, Dublin, Ireland
Andy Way away@computing.dcu.ie
CNGL, School of Computing, Dublin City University, Dublin 9, Ireland

Abstract

This paper presents a study of user-perceived vs real machine translation (MT) post-editing effort and productivity gains, focusing on two bidirectional language pairs: English—German and English—Dutch. Twenty experienced media professionals post-edited statistical MT output and also manually translated comparative texts within a production environment. The paper compares the actual post-editing time against the users' perception of the effort and time required to post-edit the MT output to achieve publishable quality, thus measuring real (vs perceived) productivity gains. Although for all the language pairs users perceived MT post-editing to be slower, in fact it proved to be a faster option than manual translation for two translation directions out of four, i.e. for Dutch→English, and (marginally) for English→German. For further objective scrutiny, the paper also checks the correlation of three state-of-the-art automatic MT evaluation metrics (BLEU, METEOR and TER) with the actual post-editing time.

1 Introduction

Machine translation (MT) has developed considerably in the last few years, to the point that it has started to be implemented in industrial translation production scenarios (DePalma, 2011). The industry is embracing MT for certain use-cases, mainly because post-editing (PE) MT output (as opposed to translating from scratch) can lead to productivity gains, particularly within well-defined technical domains (cf. Plitt and Masselot, 2010). However, on the whole, sceptical attitudes remain towards the real benefits of implementing workflows involving MT followed by PE in an effort to speed up the translation process. The main motivation behind this study lies in the need to investigate on the basis of solid evidence the attitudes towards MT PE viz. the actual benefits it yields to obtain high-quality translations. More specifically, the paper examines the extent to which perception matches reality when comparing full MT PE with manual translation from scratch by looking at subjective evaluations (i.e. user perception of time investment, effort and preferred working method) against objective measurements (i.e. the actual time gains).

* Work done while at CNGL, School of Computing, Dublin City University, Ireland.

The rest of the paper is structured as follows. Section 2 examines related work, showing the growing attention that PE has received in the last few years, especially with regard to studies focusing on PE effort and productivity gains; Section 3 outlines the experimental design and set-up of our study, describing the methodology and materials; Section 4 presents the results obtained from the questionnaire that was given to the participants in the study, and Section 5 analyzes the key experimental results concerning real vs perceived productivity gains with PE, also in relation to three state-of-the-art automatic MT evaluation metrics. Finally, Section 6 draws some conclusions, briefly summarizing the key findings of the study.

2 Related Work

Plitt and Masselot (2010) report a productivity test conducted on MT followed by PE as compared to traditional human translation in an industrial environment. They observed that MT helped translators to substantially improve their productivity: MT followed by PE improved throughput on average by 74%, which in effect reduced translation time by 43%. Zhechev (2012) carries out a PE productivity test[1] using a CAT-based PE environment at Autodesk for nine language pairs, and he also found that MT followed by PE results in substantial productivity gains over translation from scratch (ranging from 37% to 92%, depending on the language pair).

Läubli et al. (2013) report experiments carried out in a realistic translation environment and conclude that PE led to significant time gains, even when a fully functional translators' workbench is available. Tatsumi and Roturier (2010) studied the correlation between source-text characteristics and their effects on technical and temporal PE effort on a small English–Japanese dataset. They observed strong correlation between Systran's complexity and ambiguity scores and technical PE effort, and moderate correlation between the IQ score provided by Acrolinx (a widely used authoring software product) and temporal PE effort. Poulis and Kolovratnik (2012) conduct a large-scale evaluation of MT PE aimed at estimating the business benefits of using MT for the European Parliament on 5 European language pairs. They found that on an average 21.4% of the translated segments were rated excellent by the human evaluators (i.e. requiring no PE), while 25.6% of the translations were deemed good (i.e. requiring only minor PE effort). In addition, 20.8% and 32.2% of the translations were found to be average (i.e. requiring major PE) and poor (i.e. of no use), respectively.

Koponen et al. (2012) suggested using PE time as a measure of assessing the cognitive effort involved in PE. They tried to identify different types of MT errors and correlate them with the different levels of difficulty involved in fixing them, where difficulty is measured in terms of PE time. Koponen (2012) studied the relationship between cognitive and technical aspects of PE effort by comparing human scores of perceived effort necessary with the actual edits made by post-editors for cases in which the edit distance and manual scores reflecting perceived effort diverged. The results of an error analysis performed on such data are discussed in terms of the clues that they might provide about edits requiring greater or less cognitive effort compared to the technical effort involved.

Guerberof (2009) studied the effectiveness of using MT output as opposed to translation memory fuzzy matches for the purpose of post-editing in an English→Spanish translation task. She used Language Weaver's statistical MT engine and trained it on the same TM, performing both quantitative and qualitative analyses. The main result was that the productivity of the translators as well as the quality of the translation improved when post-editing MT output, compared to when processing fuzzy matches from the translation memory database.

[1] http://langtech.autodesk.com/productivity.html.

Koehn and Haddow (2009) describe Caitra, a tool that makes suggestions for sentence completion, shows word and phrase translation options, and supports PE of MT output. They report a user study carried out with the tool involving 7 translators for the English–French language pair. Among the different types of assistance offered by Caitra, users prefer the prediction of sentence completion and the options from the translation table over the other types of assistance available for post-editing MT output. To the authors' surprise, PE received the lowest scores among all the options, both in terms of enjoyment and subjective usefulness, although PE was as productive as the other types of assistance.

In an effort to extend the initial insights presented in particular by Koehn and Haddow (2009) and Koponen (2012), this paper investigates perceived vs real productivity gains brought about by post-editing MT output compared against manual translation from scratch in the relatively open – and thus particularly challenging – news-oriented domain.

3 Set-up of the Study

3.1 Methodology and Materials

Output from the CoSyne statistical MT systems (Martzoukos and Monz, 2010) was used in this experiment, and a facility was in place to track the time required by the users to post-edit MT output and to perform manual translations from scratch on texts of similar length and complexity. The texts chosen for the study were extracted from "Today in History/Kalenderblatt" and "Beeld en Geluidwiki", the public wiki sites of the two media organizations that acted as end-user partners in the CoSyne project, namely Deutsche Welle (DW) and the Netherlands Institute for Sound and Vision (NISV).[2] These two bilingual wiki sites cover news, accounts of historical events, biographies of personalities from TV and cinema and descriptions of films and TV series. The texts considered for the experiment were representative of typical source texts used during translation production, as DW and NISV were investigating ways of incorporating MT into their workflows to translate entries of these public wiki sites.

DW chose 10 texts to be translated from German into English with 399 sentences overall, and 10 texts in the opposite direction with 390 sentences. On the other hand, 15 wiki texts were chosen for Dutch to English (394 sentences in total) and 21 for English to Dutch (346 sentences) by NISV. Staff from each organization worked on the respective texts for their own language pair. The quantity of data for each translation direction was roughly similar, with approximately 6,000 words in each of the four source languages. Each sentence was translated manually from scratch and fully post-edited after MT processing by two different participants for the same translation direction. To maximise the amount of data available, each user worked on different texts, taking on the role of experimental subject (using MT followed by PE) and control group member (translating manually from scratch), in turn.

3.2 The Questionnaire

Part of the study was conducted through a preliminary MT evaluation questionnaire given to all 20 participants,[3] which was subsequently supplemented by the collection of experimental PE data. A few initial items in the questionnaire covered basic personal information concerning the age and gender of the respondents, their role in the organization and their professional experience, as well as their previous use of MT. The remaining parts focused specifically on

[2] The URLs are www.todayinhistory.de and www.beeldengeluidwiki.nl, respectively.

[3] The full questionnaire is available at www.computing.dcu.ie/~atoral/resources/questionnaire_post-editing_perception.pdf (only the answers directly relevant to the study are discussed in this paper).

the judgements of the users on the quality of the CoSyne MT output and on their perception of PE compared against manual translation from scratch for translating wiki texts.

At the beginning of the evaluation sessions, the users were informed that for the purposes of this experiment, "post-editing" meant checking the raw output provided by the MT system against the source-language input, revising and improving it as required to obtain a final target text of publishable quality. The purpose of this was to add the final revised translation to the public wiki of their respective media organization; hence, the scenario was that of full PE, aiming for optimal quality of the final revised text (Allen, 2003: 306). In addition, it should be noted that while all participants in the experiment had experience in manual translation, none of them had been specifically trained to carry out PE in a realistic professional task. This is quite different from previous studies such as Snover et al. (2006: 227), where monolingual annotators "were coached on how to minimize the edit rate". To sum up, our study focused on a scenario in which (i) the translators were not trained specifically on PE, and (ii) the objective was publishable quality, as a means of investigating the role of full PE in industrial settings, especially in terms of the perceived vs actual productivity gains.

4 Questionnaire Results

4.1 Profiles of the Participants

At the time of completing the questionnaire, the youngest DW staff member was 38 years of age, and the oldest was 59. Overall, the average age of DW staff who conducted the experiment for the English—German language pair was just over 43 years. In contrast, the NISV employees were aged between 26 and 35, and their average age was just above 31 years. In terms of gender, 7 of the 10 DW respondents were male, and the remaining 3 female. The NISV staff were evenly split between 5 men and 5 women. In total, therefore, the sample of respondents for the two language pairs consisted of 12 men and 8 women. These individuals held a variety of roles within their media organizations (e.g. journalists, editors, etc.), and all of them contributed in various capacities to the creation, development and management of multilingual content on the public wiki sites, from which the texts for this study were extracted. As part of their work, some of these subjects frequently translated content similar to that involved in the experiments conducted for this study in the same language pairs.

In terms of the experience in their organization, DW respondents had been with their employer from a minimum of 1 year (which was the case for a freelance collaborator) to a maximum of 15 years (a senior project manager), with the average being slightly more than 5 years. In contrast, the time spent by NISV staff with their current employer ranged from a minimum of 4 months (in the case of a recently hired professional) to a maximum of 5 years (the chief wiki editor), with the average being just under 3 years of continuous employment.

4.2 Evaluation for EN—DE at DW

This section concerns the questionnaire answers provided by DW staff for the English—German language pair, while Section 4.3 focuses on English—Dutch translation, evaluated by the NISV employees. It should be kept in mind in this respect that in the remainder of this analysis the evaluations for each of the four translation directions under study were formulated by 5 people, supplemented by a comparable control group of the same size.

After performing full PE on the CoSyne MT output to bring it to publishable quality, the users were asked a number of questions focusing on their subjective perception of PE. In particular, the respondents were asked which working method in their opinion involved more effort, i.e. PE of MT output or manual translation from scratch; the lower the score (on a 5-

point Likert scale), the more negative the opinion held by the respondents on the cost-effectiveness of PE compared to manual translation. As shown in Table 1, for the German—English language pair the responses tended to cluster in the lower part of the spectrum, with average scores of 2.0 for the translation direction into English, and 1.75 for translations into German.[4]

Translation direction	*MT with post-editing*				*Manual translation*	*Don't know*	***Avg. (1-5)***
DE→EN	2	2		1			***2.0***
EN→DE	2	1	1			1	***1.75***

Table 1. Effort perception: MT with post-editing vs manual translation for EN—DE.

Next, the users were asked to comment on which working method they thought was faster (thus relying on their own perception) for the two translation directions, i.e. post-editing MT output or manual translation from scratch. The answers are summarized in Table 2, and it is clear that for the English—German language pair there was a strong perception that manual translation from scratch was faster than PE.

Translation direction	*MT with post-editing*				*Manual translation*	*Don't know*	***Avg. (1-5)***
DE→EN		1		2	2		***4.0***
EN→DE			1	2	2		***4.2***

Table 2. Speed perception: MT with post-editing vs manual translation for EN—DE.

Finally, the questionnaire asked which working method the users preferred between post-editing MT and translating from scratch. Table 3 presents the answers given by DW staff.

Translation direction	*MT with post-editing*		*Manual translation*	***Avg. (1-3)***
DE→EN	2	1	2	***2.0***
EN→DE	1		4	***2.6***

Table 3. Overall preference: MT with post-editing vs manual translation for EN—DE.

In the case of German→English translations, there is a neutral situation, with the average score being 2 out of a 3-point scale: 1 respondent had no preference (middle column), while the other two pairs of participants expressed opposite opinions. This might suggest that personal predisposition and possibly expectations related to MT quality could be playing a role in this area. Interestingly, in the opposite direction (English→German) there is a marked preference for manual translation from scratch, with an average score of 2.6 out of 3. However, this opinion was not unanimous, because 1 out of the 5 interviewees stated that they preferred post-editing MT output rather than translating manually from scratch, again pointing to the role of subjective variability in this area.

An important point must be added in this respect, which also applies to the EN—NL language pair, analyzed in Section 4.3, namely that the best human translators are not necessarily the best post-editors. This is particularly relevant here, given that our experimental sub-

[4] In Tables 1-6, for each language pair, the figures in the right-most "Avg." column indicate the average scores (out of 5 or out of 3, as shown). The integer numbers in the remaining columns show how many of the 5 evaluators for that language pair provided the relevant response between the two extremes (the columns in between corresponding to intermediate values along the Likert scales, with the middle one representing neutral "neither one, nor the other" answers). Empty cells mean that no respondents provided that answer, while the "Don't know" column records the number of respondents who did not have a clear answer on the specific point.

jects had not received any specific training in PE, and were naive to the task, while they had varying levels of experience in traditional translation; similarly, there is no reason to suggest that a negative opinion, or a preconceived dislike, of PE leads to a poor PE performance, e.g. one that is slower than human translation from scratch. Due to lack of space, in this paper we do not investigate the relationship between these dimensions concerning the perception and the reality of PE productivity gains, but these are issues that deserve further study.

4.3 Evaluation for EN—NL at NISV

This section concerns the questionnaire results for English—Dutch, and follows the same structure of Section 4.2 for ease of comparison with the results analyzed for the English—German language pair. With regard to the perceived effort that the NISV participants associated with using MT output followed by PE to translate wiki entries, as opposed to manual translation from scratch, Table 4 shows the results for translations from Dutch, where the overall score is slightly in favour of MT with PE (2.75 out of 5 points, with one "don't know" answer); however, the opposite is true for translations into Dutch, for which the users clearly attribute much more effort to MT followed by PE.

Translation direction	*MT with post-editing*				*Manual translation*	*Don't know*	***Avg. (1-5)***
NL→EN		2	1	1		1	***2.75***
EN→NL	4	1					***1.2***

Table 4. Effort perception: MT with post-editing vs manual translation for EN—NL.

The NISV employees were also asked which of the two working methods they perceived to be faster, and the answers to this question are summarized in Table 5.

Translation direction	*MT with post-editing*				*Manual translation*	*Don't know*	***Avg. (1-5)***
NL→EN	1	2		1		1	***2.25***
EN→NL				1	3	1	***4.75***

Table 5. Speed perception: MT with post-editing vs manual translation for EN—NL.

There is a slight preference for using MT followed by PE in the NL→EN translation direction: the average score in that case is 2.25 out of a 5-point scale. However, the opposite is true in the other translation direction, with a total score of 4.75 out of 5.0, and again 1 respondent who opted for "don't know". This means that in general for translations into Dutch, manual translation from scratch was thought to be much less time-consuming than post-editing MT output.

Finally, NISV users were asked about their overall preference between post-editing MT output on the one hand and manual translation from scratch on the other to translate wiki texts between Dutch and English, and Table 6 shows the answers in this respect.

Translation direction	*MT with post-editing*		*Manual translation*	***Avg. (1-3)***
NL→EN		1	4	***2.8***
EN→NL		1	4	***2.8***

Table 6. Overall preference: MT with post-editing vs manual translation for EN—NL.

For both translation directions there was the same overall score, clearly in favour of manual translation from scratch, i.e. 2.8 on a 3-point scale. The strong tendency to indicate manual translation as the preferred working method for English—Dutch seems to be less influenced

by personal inclinations, with 1 respondent having no preference in both cases (recorded in the middle column), but all the other members of the sample favouring manual translation.

5 Analysis of Experimental Results

Following on from the subjective evaluation presented in Section 4, this section focuses on the objective experimental results. In particular, in Section 5.1 we zoom in on the actual user performance, comparing real versus perceived PE time and productivity gains for all language pairs. To add a further objective dimension to this part of the study, in Section 5.2 we calculate the correlation between actual PE time and the scores of three state-of-the-art automatic evaluation metrics for the MT output of all four translation directions.

5.1 Time Tracking: Real vs Perceived Productivity Gains with PE

During the evaluation sessions, timestamps were recorded for the manual translation from scratch of the documents as well as for MT PE; we can therefore compare the time taken when translating from scratch with the time spent post-editing MT output, normalizing the measure to the average per single word on the source side. These measures allow us to objectively quantify any productivity gains that are achieved with PE for each translation direction; this can, in turn, be compared with the users' perceptions of time gains.

Following Plitt and Masselot (2010) and Zhechev (2014: 9), we filtered out from our analysis the data related to the texts for which the translation took substantially more time than the others, on the basis of the average processing time per source-language word. We decided to discard the texts for which the processing time (either translation from scratch or PE) took more than 7 seconds per word, which were considered outliers for our purposes. This corresponds to only 2 texts, both for the German→English translation direction, accounting for 3.6% of the overall data sets, but which consumed a notably higher proportion (9.1%) of the overall translation time; we therefore did not want these outliers to unduly skew the results.

Figure 1 shows the average time taken to translate the documents for each translation direction (measured in seconds per word, calculated on the source language), both when translating from scratch (columns HT) and when post-editing MT output (columns PEMT).

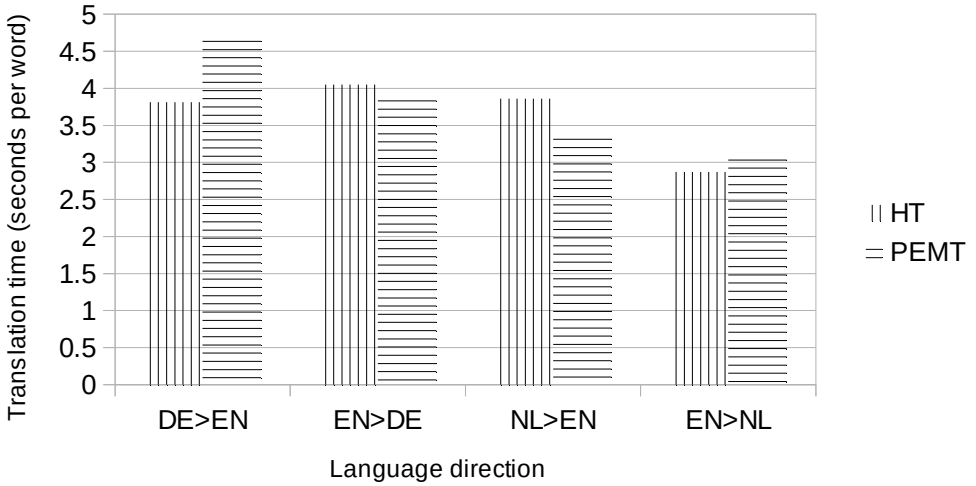


Figure 1. Manual translation and PE time (HT=translation from scratch; PEMT=post-editing).

The results of this part of the objective evaluation are mixed. For English→German and Dutch→English, post-editors took less time than the professionals translating from scratch on average (3.5% and 13.6% productivity gains, respectively). While 13.6% certainly is an encouraging figure, 3.5% represents a modest productivity gain to justify the investment in MT. In real translation workflows, productivity gains of only a few percentage points thanks to PE would be regarded as negative results, in the sense that the management of translation agencies or multilingual departments of large companies would be reluctant to introduce MT with such low gains for a particular language pair; however, it should be noted that our users had no training or experience in PE, and that even relatively marginal productivity gains of this kind would correspond to potentially significant savings across multiple language pairs, such as those typically covered by large multinational companies. In addition, it seems reasonable to expect PE-related productivity gains to rise as staff receive training and acquire experience in the task for a specific language pair. In contrast, for the other two translation directions (i.e. German→English and English→Dutch), post-editing MT output took more time than translating from scratch, leading to productivity losses of 19.16% and 7.88%, respectively, which again can be attributed, at least to some extent, to the fact that the participants in our study had no prior training in PE.

Next, we compared these results with the users' judgements, i.e. their perception regarding both effort and speed, as well as their favourite working method (cf. Tables 1, 2 and 3 for EN—DE, and Tables 4, 5 and 6 for EN—NL). Figure 2 shows these judgements on a 5-point scale.[5] The closer the value of a judgement is to 5, the stronger the preference given to manual translation from scratch. Conversely, the closer the value is to 1, the stronger the preference for post-editing MT output.

Figure 2 also includes the PE gains in terms of time (based on the results presented in Figure 1). The PE time gains are scaled up to a 5-point scale with the following formula:

```
If (PE time gain < 0%)
        Time gain = 3 - 2 * abs(PE time gain)
else
        Time gain = 3 + 2 * 2 * abs(PE time gain)
```

The equation evaluates to 3 (middle value on the 5-point scale, corresponding to no winner between PE and manual translation from scratch) if the PE gain is 0%. The score equals 5 (highest score, i.e. maximum preference for translation from scratch) if PE takes double the time than translating from scratch (i.e. PE productivity gain -50%). Finally, the score equals 1 (lowest score, i.e. maximum advantage for PE) if PE takes half the time compared to translating from scratch (i.e. PE productivity gain 100%).

Analyzing the results shown in Figure 2, we obtain three main findings. First of all, when comparing the subjective judgments with the actual time gains, we notice that most of the judgments are biased towards translation from scratch, with the only exceptions being speed for NL→EN and favourite method for DE→EN. When considering all four translation directions, the scores given to effort, speed and favourite working method are on average 1.07, 0.79 and 1.24 points higher, respectively, than the actual time gain score.

[5] While perceptions of PE effort and speed (as opposed to manual translation from scratch) were originally expressed on a 1-5 scale, overall preference was scored on a 1-3 scale. In the interest of consistency, the values for preference are thus scaled up to a 1-5 scale.

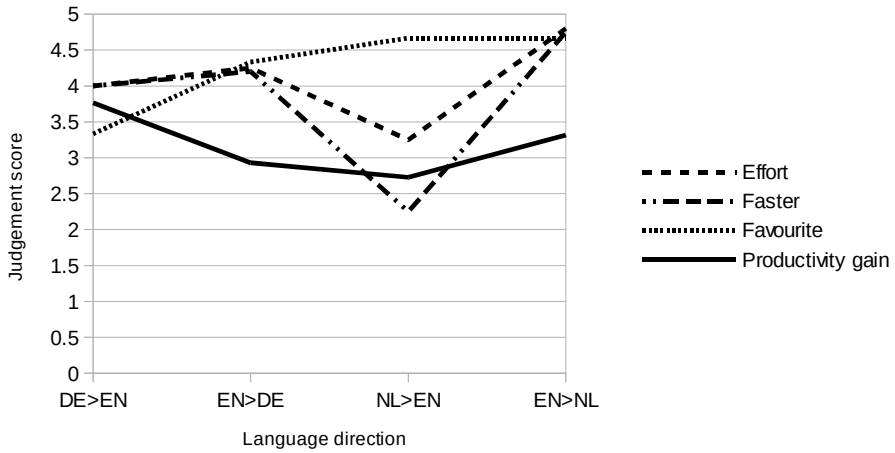


Figure 2. Comparison of users' perceptions and real time gain with PE.

Secondly, all the user judgments tend to express consistent preference for manual translation from scratch over PE of MT output, especially for EN→DE and for EN→NL. The exception to this is NL→EN, where the users perceived PE to be slightly faster (2.25) but requiring more effort than translating from scratch (3.25), but then expressed an overwhelming preference for manual translation as their favourite method (4.67).

Thirdly, while the judgments regarding speed and effort vary across translation directions, probably reflecting the differences in time gain, the results for the favourite working method are rather stable across translation directions, and are particularly high, with similar scores indicating a strong preference for manual translation from scratch, for three translation directions out of four (i.e. all of them except DE→EN). We can thus conclude that the users' overall preference for translation from scratch as their favourite working method is independent of the actual time gain/loss and from real productivity advantages when comparing translation from scratch with MT PE.

5.2 Automatic Evaluation Metrics and PE Gains

The data obtained in this experiment, which consisted of human translations from scratch and of post-edited MT output for the four translation directions, can be considered as alternative sets of reference translations, since the post-edited MT output is of publishable quality; the main difference between the two sets of references is that the translations from scratch were created independently of the MT system, while the post-edited versions were based initially upon the raw statistical MT output, with subsequent revision. One interesting observation in this respect is that, while MT output for a given language pair is consistent (the strengths and weaknesses of a system remain stable, and thus no substantial qualitative variation occurs in the output), human translators as well as post-editors cannot be assumed to be consistent.

This is due only in part to individual differences and idiosyncracies, as two human translators may prefer different, but equally good, translations of the same source passage simply because of personal stylistic preference. One must also consider the inherent variability of human behaviour, including when the same person does a relatively repetitive translation over a period of time: even though they may come across identical phrases at different points (as was likely e.g. with the biographies included in our data sets), they might, more or less consciously, end up translating them differently. This variable behaviour applies even more to post-editors: the degree and the type of corrections made by the same as well as by different individuals to the MT output for one language pair are likely to be unpredictably inconsistent.

We thus evaluated the raw MT output against both the human translations and the post-edited MT output using three state-of-the-art automatic evaluation metrics, namely BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006). The automatic MT evaluation scores are shown in Figure 3. Following the conventions used in Snover et al. (2006), the scores against references translated from scratch are named after the metric (i.e. BLEU, METEOR and TER), while the scores against the post-edited references are named appending the prefix H (i.e. HBLEU, HMETEOR and HTER, respectively).

If we compare the human translation scores against the PE scores, we can see that the PE scores are consistently better than the human translation scores for all the translation directions across all the metrics (bearing in mind that TER is an error rate metric, so unlike the other two metrics, the lower the score the better). This is expected, as the automatic metrics rely on *n*-gram matching of surface forms.[6] Hence, a reference translation that is based on the output of the MT system is likely to have a higher overlap with the raw MT output than a reference that is created independently of the MT output.

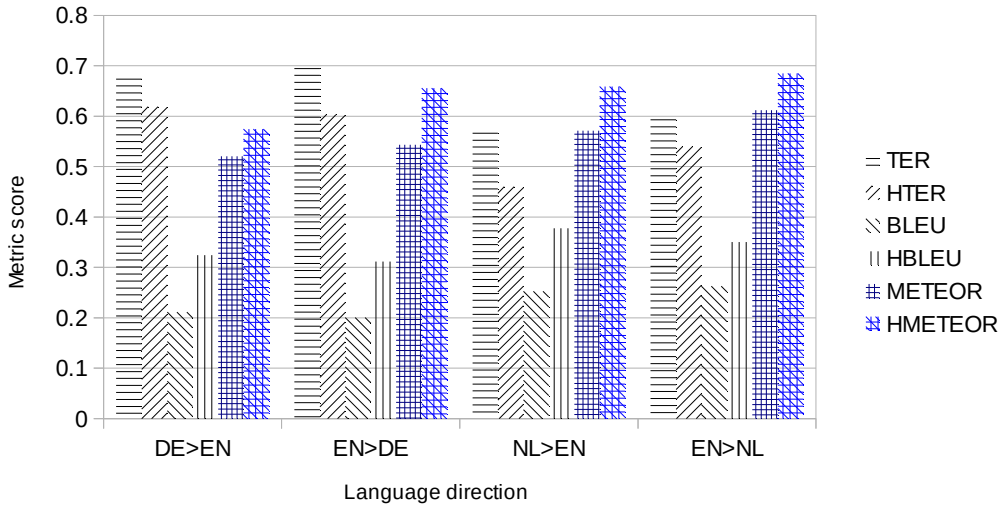


Figure 3. Scores of automatic evaluation metrics.

We then explore whether the scores obtained via the automatic evaluation metrics correlate with translation time, both when translating from scratch and when post-editing. For each translation direction we compute the Pearson correlation between the translation times and the scores of each automatic metric obtained on the MT output at the document level. A limitation that should be taken into account when interpreting any findings extracted from the data relates to the small size of the samples (the number of documents per translation direction varies from 10 to 21).

Intuitively, one would expect translation times to correlate with the scores of the automatic metrics (i.e. the better the score, the lower the PE time). Thus we would expect negative correlations between translation times and BLEU and METEOR scores (i.e. the longer the translation time the worse the metric score) and positive correlations between translation times and TER scores (i.e. the longer it takes to translate, the higher the error rate expressed by TER).

[6] METEOR also considers additional linguistic information, such as stems and synonyms.

It should also be noted that all correlations are calculated with respect to the reference that was translated from scratch. We adopted this approach as these references are independent of the raw statistical MT output, while post-edited references would be biased towards the MT system. Table 7 shows the correlations, which are presented for each translation mode (PEMT, HT), for each metric (TER, BLEU, METEOR) and for each translation direction, plus for all the translation directions combined.[7] The correlations are shown in bold (expected ones), in italics (unexpected ones), and in normal font (no correlation). We consider there to be no correlation if the value is between -0.2 and 0.2.

Translation	*TER*		*BLEU*		*METEOR*	
direction	*PEMT*	*HT*	*PEMT*	*HT*	*PEMT*	*HT*
DE→EN	**0.79**	**0.41**	**-0.93**	**-0.24**	**-0.94**	-0.10
EN→DE	*-0.58*	0.20	*0.53*	**-0.28**	*0.41*	**-0.24**
NL→EN	-0.20	**0.45**	0.00	**-0.38**	**-0.23**	**-0.23**
EN→NL	0.00	-0.05	-0.03	0.19	0.01	0.19
All	**0.25**	**0.28**	**-0.28**	**-0.23**	**-0.42**	**-0.25**

Table 7. Correlations between translation time and automatic metrics.

Considering each of the four translation directions separately, translations from scratch (columns HT) seem to correlate more consistently (out of 12 results, there are 7 expected correlations, 5 no correlations and no unexpected correlations) than post-edited translations (columns PEMT), for which the picture is rather mixed (4 expected correlations, 5 no correlations and 3 unexpected ones). Aggregating the data for all the translation directions, we observe consistent results regardless of the metric (TER, BLEU and METEOR) or the translation method (PEMT, HT): all the correlations are as expected, their values ranging from ±0.23 to ±0.42.

6 Conclusions

We have presented a study of real vs perceived PE productivity gains for the German—English and Dutch—English bidirectional language pairs. Previous studies such as Plitt and Masselot (2010) and Zhechev (2012) had looked at PE productivity gains compared to manual translation. However, in a similar vein to Koehn and Haddow (2009) and Koponen (2012), this study has crucially brought into the picture the perceptions of the users in terms of PE effort and speed, comparing them to the actual PE time gains.

We have found a bias in favour of translation from scratch across all four translation directions for all the levels of perception considered (speed, effort and favourite working method). While the perception of speed and effort seems to correspond to the actual gains to some extent, the favourite working method remains independent of the time gain and is consistently in favour of manual translation from scratch, thus pointing to a lingering sceptical attitude towards the benefits of PE, regardless of actual productivity gains. We have found these for Dutch→English and, albeit more modestly, for English→German, while PE led to productivity losses over manual translation from scratch for English→Dutch and German→English; crucially, PE was consistently the least preferred working method compared to translation from scratch, regardless of the productivity gains or losses.

In addition, we have explored the correlations of three standard automatic evaluation metrics (BLEU, METEOR and TER) with translation time, both when translating manually and when post-editing MT output. Although we can only reach tentative conclusions due to the limited data analyzed across the four language pairs, both manual translation and post-

[7] Note that these correlations are calculated by aggregating the data across all the four translation directions. Thus any findings drawn may be limited due to lack of cohesion of the data.

editing lead to weak correlations between the time to complete the task and the scores of the automatic evaluation metrics.

Acknowledgements

This work was partially funded by the European Commission through the CoSyne project (Grant FP7-ICT-4-248531) and partially supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of CNGL at Dublin City University. The authors would like to thank Tim Koch and Peggy van der Kreeft at Deutsche Welle and Jaap Blom and Johan Oomen at the Netherlands Institute for Sound and Vision, who ran the field tests on which this study was based, and the 20 professionals at both DW and NISV who took part in them. Thanks are also due to the anonymous referees who commented on an earlier version of this paper.

References

Allen, J. (2003). Post-editing. In Somers, H. (ed) *Computers and Translation: A Translator's Guide*. John Benjamins. 297-317.

Banerjee, S. and Lavie, A. (2005). An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL-2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*. Ann Arbor, Michigan. 65-72.

DePalma, D.A. (2011). *The Market for MT Post-Editing in 2011*. Cambridge, MA: Common Sense Advisory. Available at www.commonsenseadvisory.com/AbstractView.aspx?ArticleID=2202.

Guerberof, A. (2009). Productivity and quality in MT post-editing. In *Proceedings of the MT Summit XII Workshop on "Beyond Translation Memories: New Tools for Translators MT"*. Ottawa, Canada. 8 pages.

Koehn, P. and Haddow, B. (2009). Interactive assistance to human translators using statistical machine translation methods. In *Proceedings of the MT Summit XII: Twelfth Machine Translation Summit*. Ottawa, Ontario, Canada. 73-80.

Koponen, M. (2012). Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the WMT 2012 7th Workshop on Statistical Machine Translation*. Montréal, Canada. 181-190.

Koponen, M., Aziz, W., Ramos, L. and Specia, L. (2012). Post-editing time as a measure of cognitive effort. In *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice*. San Diego. 11-20.

Läubli, S., Fishel, M., Massey, G., Ehrensberger-Dow, M. and Volk, M. (2013). Assessing post-editing efficiency in a realistic translation environment. In *Proceedings of the MT Summit XIV Workshop on Post-editing Technology and Practice*. Nice, France. 83-91.

Martzoukos, S. and Monz, C. (2010). The UvA system description for IWSLT 2010. In *Proceedings of the 7th International Workshop on Spoken Language Translation*. Paris, France. 205-208.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA. 311-318.

Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localization context. *Prague Bulletin of Mathematical Linguistics*, 93:7-16.

Poulis, A. and Kolovratnik, D. (2012). To post-edit or not to post-edit? Estimating the benefits of MT post-editing for a European organization. In *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice*. San Diego. 60-68.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas "Visions for the Future of Machine Translation"*. Cambridge, MA. 223-231.

Tatsumi, M. and Roturier, J. (2010). Source text characteristics and technical and temporal post-editing effort: what is their relationship? In *Proceedings of the JEC 2010 Second joint EM+/CNGL Workshop "Bringing MT to the user: research on integrating MT in the translation industry"*. Denver, Colorado. 43-51.

Zhechev, V. (2012). Machine translation infrastructure and post-editing performance at Autodesk. In *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice*. San Diego. 87-96.

Zhechev, V. (2014). Analysing the Post-Editing of Machine Translation at Autodesk. In Balling, L.W., Carl, M., Simard, M., Specia, L. and O'Brien, S. (eds) *Post-Editing of Machine Translation: Processes and Applications*. Newcastle upon Tyne: Cambridge Scholars Publishing. 2-23.

Cognitive Demand and Cognitive Effort in Post-Editing

Isabel Lacruz ilacruz@kent.edu
Institute for Applied Linguistics, Kent State University, Kent OH 44240, U.S.A.

Michael Denkowski mdenkows@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA 15213, U.S.A.

Alon Lavie alavie@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA 15213, U.S.A.

Abstract

The pause to word ratio, the number of pauses per word in a post-edited MT segment, is an indicator of cognitive effort in post-editing (Lacruz and Shreve, 2014). We investigate how low the pause threshold can reasonably be taken, and we propose that 300 ms is a good choice, as pioneered by Schilperoord (1996). We then seek to identify a good measure of the cognitive *demand* imposed by MT output on the post-editor, as opposed to the cognitive *effort* actually exerted by the post-editor during post-editing. Measuring cognitive demand is closely related to measuring MT utility, the MT quality as perceived by the post-editor. HTER, an extrinsic edit to word ratio that does not necessarily correspond to actual edits per word performed by the post-editor, is a well-established measure of MT quality, but it does not comprehensively capture cognitive demand (Koponen, 2012). We investigate *intrinsic* measures of MT quality, and so of cognitive demand, through edited-error to word metrics. We find that the transfer-error to word ratio predicts cognitive effort better than mechanical-error to word ratio (Koby and Champe, 2013). We identify specific categories of cognitively challenging MT errors whose error to word ratios correlate well with cognitive effort.

1 Introduction

The task of the post-editor is to render machine translation output in a readily usable form in the target language. Anyone who has successfully struggled with strangely worded assembly instructions can attest that it is sometimes possible for a person with no knowledge of the source language to make good sense of an imperfect machine translation – provided there is sufficient context. However, most post-editing is carried out by professional translators.

Nevertheless, formal training in post-editing has only recently been introduced as a necessary part of translation training (e.g., O'Brien, 2002), and more work remains to be done to identify the critical competences that predict good post-editing performance (e.g. Almeida, 2013). Post-editing, the task of editing MT output in the target language while referring as needed to the source text in a different language, is very different from the task of translating directly from one language to another. Consequently, the cognitive processes involved in these two tasks will also differ. As a result, traditional translator training may not be ideal preparation for work as a post-editor. While translation process research has made considerable progress in recent years (see, for example, Muñoz Martín, 2014) our understanding of the post-editing process is more limited. It is now becoming important to gain a deeper understanding of the post-editing process, not only as an academic pursuit, but also as a tool to aid in the development of effective training for future translators who will work at least partly as post-editors.

Our objective in this paper is to contribute to post-editing process research by gaining more insight into effective measures of the cognitive demand an MT text imposes on the post-editor, and by investigating how that

demand relates to the cognitive effort expended by the post-editor. While this is important to understand from the theoretical and applied perspectives of post-editing process research, it is also relevant to the MT community. The amount of effort post-editors need to exert affects their productivity levels. Accordingly, a good understanding of what features of a machine translation result in higher post-editing effort levels will provide a valuable resource for machine translation researchers as they work to increase the utility of their systems. This is a different, but perhaps more effective focus than the traditional emphasis on improving adequacy compared to gold-standard reference translations (Denkowski and Lavie, 2012a).

Types of effort: Krings (2001) made significant early contributions to the study of effort in post-editing. He created a three-way categorization of different types of effort (temporal: time spent; cognitive: mental processin; and technical: physical action) and proposed that the combination of cognitive and technical effort gives rise to temporal effort. However, it is too simplistic to think that the time spent thinking without obvious action plus the time spent on keyboarding and mouse actions is the total time spent on the post-editing task. In particular, post-editors will be thinking as they type. Sometimes that thinking will not affect their typing, sometimes it will cause them to slow down slightly, and sometimes it will lead them to stop for a while. So, while it is possible to measure temporal and technical effort directly, the only options for assessing cognitive effort are indirect measures.

Technical effort, the effort required for the keyboarding and mouse actions made while editing MT output, can be measured using logging software. The software can classify, count, and time the post-editor's actions, including mouse clicks, insertions, deletions, substitutions, and shifts.

Measures of MT quality: From the utility-focused perspective of any individual post-editor, MT quality is highest when the effort required for post-editing carried out by that post-editor is least. Initially, MT quality was measured through subjective human judgments (King, 1996). It is important to note that human judgments are a measure of MT quality that is *extrinsic* to the post-editing process, since they are not made during the course of the post-editing process. They are the product of reflection and do not necessarily capture the complexities of subconscious processing during post-editing. Subsequently, a variety of automatic metrics - including TER (Snover et al., 2006), BLEU (Papineni et al., 2002), and METEOR (Lavie and Denkowski, 2009) – were developed to assess MT quality by measuring how well MT output matches one of a set of reference translations. Versions of these metrics (HTER, HBLEU, and HMETEOR) measure how well MT output matches the single post-edited version produced by an individual post-editor. Snover et al. (2006) report good correlations of the order of .6 between human judgments and each of HTER, HBLEU, and HMETEOR. These metrics are also extrinsic to the post-editing process. They do not measure the steps that were actually carried out by the post-editor. Instead, they measure the most efficient path from the MT output to the final post-edited product.

HTER can be viewed as a measure of *required* technical effort, rather than a measure of *actual* technical effort. It is computed as the ratio

$$\text{HTER} = \frac{\text{\# of required edits}}{\text{\# of reference words}},$$

where the number of edits refers to the least number of insertions, deletions, substitutions, and shifts required to convert the MT output to the final post-edited version, and the number of reference words is the number of words in the MT output. When the required technical effort for post-editing is low, HTER is also low, and MT quality is inferred to be high.

However, as observed for example by Koponen et al. (2012), HTER is not a perfect measure of actual technical effort exerted by the post editor. HTER measures the shortest route to the final product, but post-editors will often take a route that is not optimal. A simple example is where the post-editor begins to make a change in the MT output, but then reverses course and accepts the MT output without modification. The corresponding HTER will be zero. Nevertheless, the changes begun but then undone by the post-editor certainly constitute non-zero technical effort. Along with this technical effort, the post-editor will also have made cognitive effort through evaluating how to change the MT output and then deciding to abandon the change mid-stream. HTER also fails to fully capture cognitive effort (Koponen, 2012).

Pauses and cognitive effort: Overall processing rate is of great concern to businesses and translation professionals, and there are several promising studies that relate this to cognitive effort. See, for example, O'Brien (2011) and Koponen et. al. (2012). However, there are other parameters that also appear to give good insight into levels of cognitive effort during post-editing.

Previous work (Lacruz et al., 2012; Lacruz and Shreve, 2014; Lacruz and Muñoz, 2014; Green et al., 2013) has provided evidence that pauses in post-editing are indicators of cognitive effort, just as they are in other types of language production. Indeed, triangulation between keystroke logs and eye tracking data on fixations and gaze

duration demonstrate that pauses are associated with cognitive effort in monolingual language production (e.g., Schilperoord, 1996) and in translation and interpreting (e.g. Krings, 2001, Dragsted and Hansen, 2008, Shreve et al., 2011; Timarová et al., 2011). In post-editing, there is evidence that cognitively challenging edits give rise to clusters of short, possibly monitoring pauses (e.g. Lacruz et al., 2012). This motivated the consideration of Average Pause Ratio,

$$\text{APR} = \frac{\text{average time per pause}}{\text{average time per word}},$$

and Pause to Word Ratio,

$$\text{PWR} = \frac{\text{number of pauses}}{\text{number of words}}.$$

Lacruz and Shreve (2014) showed that both APR and PWR correlate well with cognitive effort, identified through detailed examination of keystroke logs. In particular, low APR and high PWR are associated with high levels of cognitive effort. This is consistent with clustering of short pauses in cognitively challenging segments. Since a short pause is not a major contributor to total pause time, the association between short pauses and cognitive difficulty also explains why O'Brien (2006) did not find an association between post-editing difficulty predicted by negative translatability indicators and Pause Ratio,

$$\text{PR} = \frac{\text{total pause time}}{\text{total time in segment}}.$$

From now, we will focus on the Pause to Word Ratio (PWR) as a measure of cognitive effort in post-editing.

2 Rationale

One issue that has not been investigated systematically is the question of what is an appropriate minimum threshold for pause length. See Green et al. (2013) for a recent discussion. It is apparent from Lacruz et al. (2012) and Lacruz and Shreve (2014) that pauses shorter than the frequently used 1000 ms or 2000 ms thresholds are important indicators of cognitive effort. However, the 500 ms threshold used in these papers was chosen somewhat arbitrarily. Others, such as Schilperoord (1996) and Green et al. (2013) have used an even lower 300 ms threshold. Keystroke logs show that a threshold below about 200 ms is not appropriate, since the time needed to routinely type consecutive characters is often in a range up to 150 ms or even more.

As the pause threshold decreases, the number of pauses will increase, and so PWR will also increase. A strong correlation between PWRs corresponding to different minimum pause thresholds would indicate that the integrity of PWR as a metric for cognitive effort during post-editing would not be compromised by some variability in the minimum pause thresholds. To investigate this issue,

- we will compute correlations between PWR values using 200 ms, 300 ms, 400 ms, and 500 ms minimum thresholds for pauses during post-editing.

The investigation of how MT quality correlates with cognitive effort in post-editing is quite recent. See, for example, Koponen (2012) and Koponen et al. (2012). In this paper, we will take a different approach by studying how well human judgments of MT quality and the automatic MT quality metric HTER, which are measures that are extrinsic to the actual post-editing process, correlate with PWR, a cognitive effort metric based on actual measurements of the post-editing process, and so intrinsic to the post-editing process. Although HTER is designed to estimate MT quality by quantifying the necessary technical effort to convert MT text into the post-edited version, it is likely that in many situations technical effort and cognitive effort will be related. Accordingly,

- we predict that increases in HTER (decreases in MT quality) will be associated with increases in PWR (increases in cognitive effort.)

The formal similarity, by which HTER measures optimal edits per word, while PWR measures actual pauses per word, and the expectation that the number of pauses will increase as the number of edits increases combine to reinforce the prediction of a positive correlation between HTER and PWR. However, since HTER estimates MT quality by measuring the distance between the MT output and the final post-edited product, but without taking into account the specific edit actions of the post-editor, there is no *a priori* guarantee that there will be a strong correlation between HTER and PWR. On the other hand, if a post-editor is asked to rate MT quality by judging how difficult an MT segment was to post-edit, their memory of actual edit actions will likely influence their judgment. It is plausible that segments that post-editors judge to be difficult will have required expenditure of higher levels of cognitive effort during the post-editing process. Accordingly,

- we predict that improvements in human MT quality ratings will be associated with decreases in PWR (decreases in cognitive effort in post-editing.)

As Snover et al. (2006) found that human ratings of MT quality correlate strongly with HTER, our two predictions are consistent with each other.

Cognitive demand imposed by MT output: The discussion so far has centered on how to measure cognitive effort expended by the post-editor in producing the final post-edited version; and how to relate that effort to extrinsic measures of the quality of the machine translation text. We have focused on PWR as a measure of cognitive effort in producing the final post-edited version. The main extrinsic measures of MT quality we discussed were HTER and human quality judgments; these did not rely on identifying specific features of the MT text.

Ultimately, we are interested in identifying intrinsic features of the source text that are associated with high levels of cognitive effort expended by the post-editor. In other words, we wish to determine which features of the source text are likely to give rise to MT output that imposes high levels of cognitive demand on the post-editor. This is a complex question, and it seems prudent to approach it step by step.

Post-editing involves three stable texts, the source text, the machine translation, and the final post-edited version. We begin by asking what intrinsic features of the machine translation, the text in the middle, place high levels of cognitive demand on the post-editor and so are associated with elevated cognitive effort on the part of the post-editor. Thus, we seek to determine *intrinsic* measures of MT quality such that increases in MT quality are associated with reductions in cognitive effort in post-editing, as measured by PWR.

This agenda was advocated by Lacruz and Muñoz (2014). Drawing on the work of Koponen (2012) and Koponen et al. (2012), they grounded their approach in the analysis of MT errors, categorized according to the linguistically based difficulty ranking proposed by Temnikova (2010) and later modified by Koponen et al. (2012). Temnikova classified MT errors into nine categories assumed to pose increasing cognitive difficulty for the post-editor. These categories are specified in Table 1.

Error ranking	Error Type
1	Correct word, incorrect form
2	Incorrect style synonym
3	Incorrect word
4	Extra word
5	Missing word
6	Idiomatic expression
7	Wrong punctuation
8	Missing punctuation
9	Word order at word level
10	Word order at phrase level

Table 1. Temnikova's MT error classification

Lacruz and Muñoz defined a cognitive demand metric for MT segments that they called Mental Load (ML). Each error in an MT segment was assigned a weight according to its type in the Temnikova classification. For

example a “Correct word, incorrect form” error was assigned weight 1; an “Idiomatic expression” error was assigned weight 6. ML for the segment was the sum of the weights for each error. Thus, a segment with three incorrect word errors, two idiomatic expression errors, and one missing punctuation error had an ML of $3\times3 + 2\times6 + 1\times8 = 29$. It was found that there was a significant strong correlation between ML and cognitive effort, as measured by Pause to Word Ratio, PWR. However, there are at least two difficulties with this analysis: Mental Load was not normalized for segment length; and Temnikova’s rankings provide order data, rather than interval or ratio data. It is not likely for example that a rank 9 error (word order at word level) is nine times more difficult to correct than a rank 1 error (correct word, incorrect form.) For the data analyzed, neither shortcoming was likely significant: the segments were mostly of very similar length, and most of the errors were low on Temnikova’s scale. In view of these two facts, it was not surprising that the total number of errors also correlated well with PWR.

In this paper, we work with a different error classification. Following the framework of the American Translators Association (ATA) grading rubric (Koby and Champe, 2013), we first classify MT errors into two categories, Mechanical (M) and Transfer (T). Mechanical errors are those that can routinely be fixed without reference to the source text. Consider, for example, an MT segment that contains the phrase *he drink the coffee*. If the machine translation is referring to a man and is consistently written in the present tense, it is clear - without reference to the source text - that this phrase contains a mechanical error and should be edited to become *he drinks the coffee.* Now consider a machine translation where the first segment is *Helen Monica helps.* This is a transfer error: without consulting the source text, it is impossible to know how to edit the segment to reflect the true meaning of the source.

- We hypothesize that the cognitive demand placed on post-editors by transfer errors is greater than the cognitive demand resulting from mechanical errors.

Error code	Error type
ILL	Illegibility
IND	Indecision, gave more than one option
MT	Mistranslation
MU	Misunderstanding of source text
A	Addition
O	Omission
T	Terminology, word choice
R	Register
F	Faithfulness
L	Literalness
FA	Faux ami
COH	Cohesion
AMB	Ambiguity
ST	Style
G	Grammar
SYN	Syntax
P	Punctuation
SP/CH	Spelling/Character
D	Diacritical marks/Accents
C	Capitalization
WF/PS	Word form/Part of speech
U	Usage

Table 2. ATA grading rubric

The American Translators Association uses a grading rubric, given in Table 2. Similarly to Angelone (2011), we construct a simplified version of the ATA rubric that we specify in Table 3. The objective is to provide a simple cognitively-based classification of MT errors that is more specific than the mechanical/transfer partition.

We combine ATA error types Mistranslation, Faux Ami, and Terminology into a single category of Mistranslation (MT); we combine error types of Addition and Omission into a single category of Omission or Addition (OA); we consider Syntax (SY) a single category; we combine error types Word Form, Grammar, and Spelling into a single category of Word Form (WF); we use a single category of Punctuation (P); and we omit the other error types related to style, since they are not relevant to the instructions given for the post-editing of machine translation output in the study in this paper. These umbrella error types do not necessarily divide cleanly into Mechanical or Transfer. In particular, Word Form errors may be either Mechanical or Transfer, depending on the context. For example, many errors of type OA will be transfer errors. However, if the text is about food and contains the phrase *fish chips*, there is no need to consult the source to realize that this should be edited to *fish and chips.*

Error code	Error type
MT	Mistranslation
OA	Omission or Addition
SY	Syntax
WF	Word Form
P	Punctuation

Table 3. Simplified error classification based on ATA rubric.

Note that we have measured cognitive effort by the intrinsic metric of pauses per word (PWR). On the other hand MT quality has been measured by extrinsic metrics, such as required edits per word (HTER). By analogy, we propose edited errors per word as a good candidate for cognitive demand, or intrinsic MT quality. Specifically, we define the Error to Word Ratio for an MT segment as

$$\text{EWR} = \frac{\#\text{ of edited errors}}{\#\text{ of words}}.$$

We will also be interested in errors of various special types. When we wish to work with errors of type X, we use the X-Error to Word Ratio,

$$\text{X-EWR} = \frac{\#\text{ of edited errors of type X}}{\#\text{ of words}}.$$

All these different EWRs can be thought of as intrinsic measures of MT quality. We investigate the extent to which it is reasonable to consider them to be measures of cognitive demand by determining how well they correlate with cognitive effort.

- We hypothesize that X-EWR will correlate more strongly with PWR when the error type X corresponds to errors that are more cognitively difficult.

In particular,

- we predict that EWR for transfer errors will correlate more strongly with PWR than will EWR for mechanical errors.

Also, in line with the general expectations of Temnikova's classification,

- we predict that EWR for transfer errors of type MT, OA, and SY will correlate more strongly with PWR than will EWR for errors of type P or WF.

3 Method

There were five participants in this study, all of whom were paid for their time. All participants had English as their first language (L1) and were highly proficient in Spanish as their second language (L2). Each participant was a student in a Master of Spanish Translation program at an American university. They had all completed a graduate level course that included instruction and practice in the process of post-editing and the use of translation memory systems as an aid in the translation process.

Source texts were extracts of Spanish language transcripts of TED talks on matters of general interest with little technical language. Four Spanish source texts were translated, each by two different adaptive machine translation systems. The adaptive MT systems learn in real time from each post-edited segment, which then impacts the translation that the MT system generates for the following segment. All participants post-edited a version of each of the four texts, two translated by one of the MT systems and two translated by the other system. Texts were divided into segments that roughly corresponded to sentences or stand-alone phrases that varied in length from 2 to 18 words, mean 9.3 words. Each participant became familiar with the set-up and procedure by post-editing a 10 segment practice text. Data corresponding to the practice text are not included in the analyses presented here. The remaining three texts, the experimental texts, contained 30 segments each. Analysis was thus carried out on 90 segments for each participant. All data was pooled since each participant post-edited potentially different MT segments. Participants post-edited the four texts in one session lasting less than two hours, although there were no time limits set for the task.

Data was collected remotely using TransCenter, a web-based translation interface that logs post-editing activity (Denkowski and Lavie, 2012b). The data used for this paper consisted of the keystroke log (for computing the number of pauses of different lengths), HTER ratings of MT quality, and user ratings of MT quality. Participants worked from their homes and were instructed to minimally post-edit. Specifically, they were asked to disregard issues of style and to focus on how well the machine translation conveyed the meaning of the source text. After participants logged in, the source segments appeared on the left of the screen and the machine translation for the first segment appeared on the right. Once the participant finished post-editing a segment, they were asked to rate that segment's suitability for post-editing on a scale from 1 to 5, as in Table 4. The scale was available for consultation at all times.

Rating	Criterion
1	Gibberish - The translation is totally incomprehensible
2	Non-usable - The translation has so many errors that it would clearly be faster to translate from scratch
3	Neutral - The translation has enough errors that it is unclear if it would be faster to edit or translate from scratch
4	Usable - The translation has some errors but is still useful for editing
5	Very good - The translation is correct or almost correct

Table 4. Criteria for user ratings of MT quality

Post-edited MT errors were classified independently by two experienced translation graders. Cases of disagreement were very limited (less than 5%). These cases were resolved through consultation between the graders.

4 Results and Discussion

Our results will be expressed in terms of correlations. We adopt Cohen's (1988) convention that a positive Pearson correlation is strong when r is at least .5, moderate when r is between .3 and .5, and weak when r is between .1 and .3. Similar conventions hold for negative r and for Spearman's ρ. We use Pearson's r for comparisons of ratio data, and Spearman's ρ for comparisons involving rank order data.

4.1 Pause Threshold

Our first objective was to assess the sensitivity of PWR to reductions in the pause threshold in 100 ms steps from 500 ms down to 200 ms. The highly significant correlations between PWR values at all of these thresholds were strong and positive, as shown in Table 5.

Pearson r	PWR-300	PWR-400	PWR-500
PWR-200	.95**	.93**	.90**
PWR-300		.98**	.96**
PWR-400			.98**

Table 5. Pearson correlations between PWRs for different pause thresholds. Significance: ** $p < .001$.

While the mean PWRs for all the pause thresholds were significantly different from each other, most differences were relatively small. However, as shown in Table 6, the difference was noticeably numerically larger for the transition from PWR-300 to PWR-200 than for the transitions from PWR-400 to PWR-300 or from PWR-500 to PWR-400. The same pattern is apparent for median values.

Center Measure	Median Value	Mean Value
PWR-200	0.50	0.71
PWR-300	0.43	0.58
PWR-400	0.40	0.50
PWR-500	0.38	0.43

Table 6. Median and mean values of PWR at different pause thresholds.

Correlations with PWR at the 200 ms threshold, while very strong and highly significant, were lower than for other comparisons. The 200 ms threshold was also dangerously close to typical typing latencies for some participants, so we took the evidence above to indicate possible contamination of pauses due to cognitive effort with pauses due to mechanical effort at this threshold. Although closer investigation would be necessary to draw firm conclusions, we chose to discard the 200 ms pause threshold for the purposes of our investigation of the relationship between utility based intrinsic measures of cognitive demand on post-editors (viewed also as a measure of MT quality) and cognitive effort in post-editing. Since the most pause information can be derived from smallest reasonable pause threshold, we will henceforth select 300 ms for the pause threshold used in computing PWR. The 300 ms choice has the benefit of conforming to some previous selections, as in Schilperoord (1996) and Green et al. (2013).

4.2 Cognitive Demand/MT Quality and Cognitive Effort

Correlations between HTER, User Ratings, and PWR: We predicted that increases in MT quality will be associated with decreases in cognitive effort. When we measure MT quality extrinsically by HTER (low HTER

corresponds to small minimal edit distance, so high MT quality) and cognitive effort by PWR, the prediction is equivalent to expecting increases in HTER to be associated with increases in PWR. This is borne out by the fact that $r = .75$, $p < .001$. In other words, as predicted, there is a highly significant strong positive correlation between PWR and HTER.

Our next prediction was a variant of the first: as MT quality improves, cognitive effort in post-editing decreases. We still measure cognitive effort by PWR, but this time we estimate MT quality by user ratings of quality – difficulty ratings made by post-editors after they complete their task. The ratings were on a scale of 1 to 5, with 1 being reserved for the most difficult segments. In these terms, decreases in user ratings were predicted to correspond to increases in PWR. This was confirmed: the Spearman correlation between user ratings and PWR was $\rho = -.71$, $p < .001$, a highly significant strong negative correlation.

We also confirmed that, as expected, there was a strong negative correlation ($\rho = -.77$, $p < .001$) between user ratings and HTER. This correlation was highly significant. Table 7 below summarizes the findings.

Correlation	HTER	User Rating
PWR-300	r = .75**	ρ = -.71**
HTER		ρ = -.77**

Table 7. Summary of correlations between HTER, User Rating, and PWR. Significance: ** $p < .001$.

Influence of Transfer and Mechanical Errors on Cognitive Effort: The next objective was to investigate how well Transfer-Error to Word Ratio (T-EWR) and Mechanical-Error to Word Ratio (M-EWR) serve as intrinsic measures of cognitive demand. Thus, in all cases, the errors considered were errors actually corrected by the post-editor.

The prediction was that transfer errors would generate more cognitive demand than mechanical errors, and so T-EWR would correlate more strongly than M-EWR with cognitive effort, measured by PWR. Likewise, since T-EWR is predicted to be a stronger intrinsic measure of cognitive demand, it should also correlate more strongly than M-EWR with extrinsic measures of MT quality, that is, extrinsic measures of cognitive demand. These predictions were confirmed by the analysis. Correlations between T-EWR and each of PWR, HTER, and User Rating were strong positive and highly significant. On the other hand, correlations between M-EWR and each of PWR, HTER, and User Rating were still highly significant but only moderate. See Table 8 for a summary.

Correlation	T-EWR	M-EWR
PWR-300	r = .56**	r = .43**
HTER	r = .60**	r = .41**
User Rating	ρ = -.61**	ρ = -.40**

Table 8. Summary of correlations of Transfer and Mechanical Error to Word Ratios with PWR, HTER, and User Rating. Significance: ** $p < .001$.

Influence of Errors in Simplified ATA Categories on Cognitive Effort: We examined correlations of Error to Word Ratios for the five error categories derived from the ATA grading rubric. In all cases, the errors considered were errors actually corrected by the post-editor. We had predicted that the more cognitively challenging error types (Mistranslation, Omission or Addition, Syntax) would be more reliable intrinsic measures of cognitive demand than Punctuation or Word Form, and so would correlate more strongly with cognitive effort (PWR) or extrinsic measures of cognitive demand (HTER or User Rating.) This was indeed the case. See the summary in Table 9 for precise details, but MT-EWR correlated strongly and very significantly and with all of PWR, HTER and User Rating, while the correlations for OA-EWR were moderate, but still highly significant. Surprisingly, correlations for SY-EWR, while highly significant, were only weak.

Other correlations were weak; those for P-EWR were highly significant, while those for WF-EWR had varied levels of significance. Accordingly, we see that EWRs for ATA categories of MT errors that were expected to be

cognitively challenging provided significant indications, albeit of variable strength, of cognitive demand in post-editing.

Correlation	PWR-300	HTER	User Rating
MT-EWR	r = .51**	r = .54**	ρ = -.58**
OA-EWR	r = .42**	r = .37**	ρ = -.39**
SY-EWR	r = .29**	r = .26**	ρ = -.28**
P-EWR	r = .17**	r = .22**	ρ = -.16**
WF-EWR	r = .05	r = .11*	ρ = -.17**

Table 9. Summary of correlations of Simplified ATA Error to Word Ratios with PWR, HTER, and User Rating. Significance: ** $p < .001$; * $p < .01$.

Influence of All Edited Errors on Cognitive Effort: However, the most reliable intrinsic measure of cognitive demand turned out to be the simple Error to Word Ratio (EWR), combining all error types. In all cases, the errors considered were errors actually corrected by the post-editor. There were strong and highly significant correlations between EWR and all of PWR ($r = .65$, $p < .001$), HTER ($r = .62$. $p < .001$), and User Rating ($\rho = -.68$, $p < .001$). This mirrors the finding in Lacruz and Muñoz (2014).

5 Conclusions and future directions

In this paper, we probed the sensitivity of the Pause to Word Ratio to changes in the pause threshold. We concluded that 300 ms is a good choice for pause threshold. It is not too short to be contaminated by normal typing activity, but is sufficiently short to capture much potentially informative pause activity.

We went on to compare PWR, an intrinsic measure of cognitive effort, with widely used metrics that have indirect relationships to cognitive effort and are often viewed as measures of MT quality. We found strong correlations between PWR and HTER, an edit to word ratio that estimates MT quality in terms of technical effort, and user ratings, that estimate MT quality in terms of perceived difficulty of post-editing.

Then we asked how we might measure cognitive demand on the post-editor. As a result of the cognitive demands placed on post-editors by features of the MT output, they must expend cognitive effort to complete the post-editing task. We chose to measure cognitive demand through Edited-Error to Word (EWR) metrics, formally analogous to the Pause to Word metric for cognitive effort and the Required-Edit to Word metric (HTER) for MT quality. Transfer errors require post-editors to review the source text to understand the meaning, while mechanical errors can reasonably be fixed without reference to the source text. The expectation is that transfer errors are more cognitively demanding to fix than are mechanical errors. This view is supported by the finding that EWR for transfer errors correlates more strongly with HTER and user ratings (MT quality measures; extrinsic measures of cognitive effort) or PWR (intrinsic measure of cognitive effort). Similarly, for other error classifications based on ATA rubrics, EWRs for those error types that were expected to be more cognitively demanding to fix correlated more strongly with PWR, HTER, and user ratings.

Results support the view that error to word ratios may be an effective way to gauge the cognitive demand imposed on post-editors by MT segments. However, these results must be viewed as preliminary, since they were generated from small samples of 90 source text segments and 5 post-editors.

Corroborating studies need to be carried out on a larger scale and supported by methodologies such as eye tracking or mouse tracking that allow direct observation of the focus of attention and have established metrics for assessing cognitive effort. It seems particularly interesting to study possible differences between the processing of transfer and mechanical errors. To gain maximum advantage, it would be worthwhile to undertake controlled experimental studies to filter out the noise of more ecological experiments. This would allow a closely focused investigation, which would potentially provide evidence to support hypotheses that could then be tested in a more natural setting.

The ultimate objective is to move beyond understanding what MT features are more or less cognitively demanding, and so require post-editors to expend more or less cognitive effort. The goal is to understand what features of the source text are associated with cognitively demanding errors in MT output. For this it may be

worthwhile to revisit the relationship between negative translatability indicators and pause data that was initiated by O'Brien (2006).

References

Almeida, Giselle. 2013. Translating the Post-Editor: An Investigation of Post-Editing Changes and Correlations with Professional Experience Across Two Romance Languages. Ph.D. Thesis, Dublin City University.

Angelone, Erik. 2010. Uncertainty, Uncertainty Management, and Metacognitive Problem Solving in the Translation Task. In Gregory M. Shreve and Erik Angelone (Eds.). *Translation and Cognition*, (pp. 17-40). Amsterdam/ Philadelphia: John Benjamins.

Denkowski, Michael and Alon Lavie. 2012a. Challenges in Predicting Machine Translation Utility for Human Post-Editors. Proceedings of the Tenth Biennial Conference of the Association for Machine Translation of the Americas, San Diego, California.

Denkowski, Michael and Alon Lavie. 2012b. TransCenter: Web-based translation research suite. In Workshop on Post-Editing Technology and Practice Demo Session, Tenth Biennial Conference of the Association for Machine Translation of the Americas, San Diego, California.

Carl, Michael, Barbara Dragsted, Jakob Elming, Daniel Hardt, and Arnt Lykke Jakobsen. 2011. The Process of Post-Editing: A Pilot Study. Proceedings of the 8th International NLPSC workshop. Special theme: Human machine interaction in translation. Copenhagen Studies in Language, 412. Frederiksberg: Samfundsliteratur.

Dragsted, Barbara and Inge Gorm Hansen. 2008. Comprehension and Production in Translation: a Pilot Study. Segmentation and the Coordination of Reading and Writing Processes. In Susanne Göpferich, Arnt Lykke Jacobsen,and Inger M. Mees (Eds.), *Looking at Eyes* (pp. 9–30). Copenhagen Studies in Language 36. Copenhagen: Samsfunds-litteratur.

Dragsted, Barbara and Inge Gorm Hansen. 2009. Exploring Translation and Interpreting Hybrids. The Case of Sight Translation. Meta: Translators' Journal, 54(3), 588-604.

Green, Spence, Jeffrey Heer, and Christopher D. Manning. 2013. The Efficacy of Human Post-Editing for Language translation. Conference on Human Factors in Computing Systems (CHI'13), April 27-May 2, 2013, Paris, France.

King, Margaret. 1996. Evaluating Natural Language Processing Systems. Communications of the Association for Computing Machinery, 29(1):73-79.

Koby. Geoffrey S. and Gertrud G. Champe. 2013. Welcome to the Ral World: Professional-Level Translator Certification. Translation & Interpreting, Vol 5(1), 156-173

Koponen, Maarit. 2012. Comparing Human Perceptions of Post-Editing Effort with Post-Editing Operations. Proceedings of the Seventh Workshop on Statistical Machine Translation (pp. 181-190), Montréal (Canada).

Koponen, Maarit, Wilker Aziz, Luciana Ramos, and Lucia Specia. 2012. Post-Editing Time as a Measure of Cognitive Effort. Proceedings of the Tenth Biennial Conference of the Association for Machine Translation of the Americas.Workshop on Post-Editing Technology and Practice (pp. 11-20), San Diego, California.

Krings, Hans P. 2001. Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes. Geoffrey S. Koby (Ed.). Kent, Ohio: Kent State University Press.

Lacruz, Isabel, Gregory M. Shreve, and Erik Angelone. 2012. Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing: A Case Study. Proceedings of the Tenth Biennial Conference of the Association for Machine Translation of the Americas. Workshop on Post-Editing Technology and Practice (pp. 29-38), San Diego, California.

Lacruz, Isabel and Gregory M. Shreve. 2014. Pauses and Cognitive Effort in Post-Editing. In Sharon O'Brien, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia (Eds.), *Post-editing of Machine Translation: Processes and Applications*. Cambridge Scholars Publishing

Lacruz, Isabel and Ricardo Muñoz Martín. 2014. Pauses and Objective Measures of Cognitive Demand in Post-Editing. Paper presented at the American Translation and Interpreting Studies Association Conference, New York, April 2014.

Lavie, Alon and Michael Denkowski. 2009. The METEOR Metric for Automatic Evaluation of Machine Translation. Machine Translation Journal, 23 (2-3), 105-115.

Muñoz Martín, Ricardo. 2014. A Blurred Snapshot of Advances in Translation Process Research. In Muñoz Martín (Ed.), *Minding Translation*, MonTi, Special Issue 1, Universidad de Alicante, Spain.

O'Brien, Sharon. 2002. Teaching Post-Editing: A proposal for course content. Proceedings of the 6th European Association for Machine Translation Workshop "Teaching Machine Translation," 14-15 November, Centre for Computational Linguistics, UMIST, Manchester, England.

O'Brien, Sharon. 2005. Methodologies for Measuring the Correlations Between Post-Editing Effort and Machine Text Translatability. Machine Translation, 19(1): 37-58.

O'Brien, Sharon. 2006. Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output. Across Languages and Cultures, 7(1), 1-21.

O'Brien, Sharon. 2011. Towards Predicting Post-Editing Productivity. Machine Translation, 25, 197-215.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

Schilperoord, Joost. 1996. It's About Time: Temporal Aspects of Cognitive Processes in Text Production. Amsterdam: Rodopi.

Shreve, Gregory M., Isabel Lacruz, and Erik Angelone. 2011. Sight translation and Speech Disfluency: Performance Analysis as a Window to Cognitive Translation Processes. In Cecilia Alvstad, Adelina Hild, and Elisabet Tiselius (Eds.), *Methods and Strategies of Process Research* (pp. 121–146), Amsterdam/ Philadelphia: John Benjamins.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Miccuilla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotations. Proceedings of Association for Machine Translation in the Americas (pp. 223-231), August 8-12, 2006, Cambridge, Massachusetts, USA.

Specia, Lucia, Nicola Cancedda, Marco Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. Proceedings of the 13th Annual Conference of the European Association for Machine Translation (pp. 28-35), Barcelona, Spain.

Temnikova, Irina. 2010. A Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment. International Conference Language Resources and Evaluation (LREC2010). Valletta, Malta, May 17-23.

Timarová, Sárka, Barbara Dragsted, and Inge Gorm Hansen. 2011. Time Lag in Translation and Interpreting. In Cecilia Alvstad, Adelina Hild, and Elisabet Tiselius (Eds.), *Methods and Strategies of Process Research* (pp. 121-146), Amsterdam/ Philadelphia: John Benjamins.

Vocabulary Accuracy of Statistical Machine Translation in the Legal Context

Jeffrey Killman jkillman@uncc.edu
Department of Languages and Culture Studies, University of North Carolina at Charlotte, USA

Abstract

This paper examines the accuracy of free online SMT output provided by Google Translate (GT) in the difficult context of legal translation. The paper analyzes English machine translations produced by GT for a large sample of Spanish legal vocabulary items that originate from a voluminous text of judgment summaries produced by the Supreme Court of Spain. Prior to this study, this same text was translated into English but without MT and it was found that the majority of the translation solutions that were chosen for the said vocabulary items could be hand-selected from mostly EU databases with versions in English and Spanish. The paper argues that MT in the legal translation context should be worthwhile if the output can consistently provide a reasonable amount of accurate translations of the types of vocabulary items translators in this context often have to do research on before being able to effectively translate them. Much of the currently available translated text used to train SMT comes from international organizations, such as the EU and the UN which often write about legal matters. Moreover, SMT can use the immediate co-text of vocabulary items as a way of attempting to identify correct translations in its database.

1. Introduction

Legal translation is often considered one of the most challenging areas of human translation practice. According to Alcaraz and Hughes, "Probably the greatest single difficulty encountered initially by legal translators is the unfamiliarity of the vocabulary characteristic of this type of discourse" (2002, p. 16), while the second major source of difficulty is the peculiarity of the morphology and syntax (2002, p. 18). Some traditional advice "is to trust nothing, to suspect everything, to check all terms in reliable dictionaries and to develop a close familiarity with the language of the law by constant and careful reading in both languages" (Alcaraz and Hughes, 2002, p. 43). Machine translation (MT) has, to say the least, typically not been recommended in the legal translation context. A perceived high risk of getting things wrong has likely continued to compel legal translators to continue to rely on more time-tested traditional approaches, while overlooking the recent technological improvements which have been made to MT, in particular statistical machine translation (SMT), and the particular aspects which might now make it a worthwhile tool in the legal translation context. This paper investigates how the output of SMT might benefit the translator of a difficult legal text. According to Forcada, "What one needs is to identify the contexts in which one can use MT effectively and to know what can be expected of it" (2010, p. 215). This paper simply seeks to identify what can be expected of MT in the legal translation context.

The specific MT system this paper looks at is the ubiquitous free online (open-domain) SMT system Google Translate (GT). Until somewhat recently GT was mostly reserved for getting the gist of a website or document in a foreign language. It is now being taken more

seriously. GT appears in some of the leading-edge technologies of translation memories (such as the newer versions of SDL Trados and Déjà Vu) to immediately provide translators with the option to post-edit its output when there are no matches in the translation memory databases themselves. It is thought that this might potentially help translators, depending on their needs or preferences. For instance, translators might save time and effort, in typing or reflecting on the translation of a piece of source text or in having to do research online or elsewhere to come up with it. Translators are not always provided a translation solution by a client or by means of a particular set of past translations or a particular translation tool (e.g. a custom-built MT system or termbase) they might be instructed to use in a professional setting. They might have to come up with solutions on their own, and so it is thought that MT output might bear some of the burden.

What if post-editing GT output at the very least could provide translators accurate translations of vocabulary items in a legal text they might have difficulty translating? This might be a modest yet indeed helpful expectation to have for the quality of the output in the context of legal translation, where one often has to spend a good deal of time doing research on such items (Cao, 2007; Biel, 2008; Monzó, 2008). True, support in the areas of morphology and syntax, the second major source of difficulty according to Alcaraz and Hughes (2002), would be most welcome too, but sentences in legal documents tend to have complex structures and most MT systems, whether rule-based or knowledge-based, handle sentences better when they are simple. SMT, in particular, has developed a good reputation for producing good word and short phrase translations, while performing not so well when it comes to overall sentence grammaticality. Moreover, much of the currently available translated text from which SMT statistically draws its translation output comes from international organizations that often write about legal matters, such as the EU and the UN (Koehn 2010, p. 53). On the basis of these premises, this paper seeks to capitalize on what SMT might be particularly good at and what human translators in the context of legal translation are particularly known for experiencing difficulty with (Alcaraz and Hughes, 2002; Cao, 2007; Biel, 2008; Monzó, 2008) and might welcome technological support with.

The paper places its focus on the accuracy of the terminological and phraseological translation choices provided in English by GT for a selection of 621 varyingly difficult vocabulary items included in a Spanish national legal text of judgment summaries: the Civil Division (Sala de lo Civil/Sala Primera) section of the Crónica de la Jurisprudencia del Tribunal Supremo: 2005-2006 (Reports of Cases before the Supreme Court: 2005-2006)[1]. The Crónica de la Jurisprudencia del Tribunal Supremo is "the work which is drafted under the same name annually by authorities (Gabinete Técnico) of the Supreme Court and intended to disseminate the judgments which for various reasons can be considered to be particularly relevant or generate wider interest" (Consejo General del Poder Judicial, n.d., translation ours).

I was involved in the translation of the Civil Division section into English in 2006-2007, which was a part of the coursework included in a legal translation certificate course at the University of Castile-La Mancha in Spain. I found in the majority of cases that consulting multilingual EU resources was especially useful in order to efficiently translate the majority of the vocabulary items that were found to be difficult. In the present study, I will compare machine translations of these items as rendered by the ubiquitous GT, to test the extent to which this popular free online SMT system might translate correctly legal vocabulary that one might otherwise have to spend time and effort on researching.

I will start by providing a brief overview of the recent research on post-editing with GT. I will then discuss in detail the vocabulary items under study, how we approached their transla-

[1] The author was involved in the translation of a thousand or so pages of both this version and the 2004-2005 version, which was published in 2006 by Spain's General Council of the Judiciary.

tion back in 2006-2007, and why SMT is likely able to translate them correctly, before providing details on the methodology. To conclude, I will discuss the results of the test and the implications of the findings.

2. Recent GT Research

According to Pym, "Recent research (Pym 2009, Garcia 2010, Lee and Liao 2011) indicates that, for Chinese-English translation and other language pairs, statistical MT [in particular GT] is now at a level where beginners and Masters-level students with minimal technological training can use it to attain productivity and quality that is comparable with fully human translation, and any gains should then increase with repeated use" (2013, p. 2). While none of these studies includes the English-Spanish language pair, "it is reasonable to assume that similar results might emerge in tests using the FIGS languages (French, Italian, German, Spanish), for which data (parallel texts, grammatical rules) have been collected over a longer period of time" (Garcia, 2010, p. 18).

Two of these three recent studies (Pym, 2009 and Lee and Liao, 2011) report on GT's vocabulary accuracy. One of the groups in Pym's study reported appreciating some of the terminology proposed by GT (p. 141). Lee and Liao (2011) include significantly more information than Pym (2009) about vocabulary accuracy in their study. In particular, they find that in the linguistically weaker of the two groups of students that were subjects of their study, "the more words from the MT text a student uses, using [the] sentence as a unit, the less likely a student would make a mistake in translating that particular sentence" (p.128). They report "that the students recognize they can use the MT directly if the meaning is intact, and they would only have to do a little tweaking" (p.128). They note a number of instances where GT translated more accurately than students the contextually appropriate meaning of an ambiguous piece of language (pp.136-137), as well as the specific meanings of adjectives and nouns found in collocations (pp.137-138). As regards the appropriate wording of concepts, they even find that "The divergence in register between the With MT and No MT students was evident" (p. 133).

These studies touch upon the potential of SMT as a viable source of vocabulary translation support. The present study seeks to determine exclusively how accurate SMT is at translating legal vocabulary translators often have to research. It compiles significant empirical data in this very important area.

A focus on lexical data is more practical than one on the morphology and syntax of machine translations of the judgment summaries contained in the Civil Division text. In the majority of cases judgment summaries are drafted in sentences that are unusually long and complex, with features such as multiple subordination and postponement of the main verb until very late in the sentence. Long, complex sentences (which are also common in many other legal text types such as statute laws, judicial rulings, and regulations) will likely not be translated correctly by SMT, or any other type of MT for that matter. What SMT basically does is weigh its decisions on the statistical patterns of various combinations of words found in a sentence and depending on the particular system, the technology might do so with or without traditional MT methods using grammatical rules (in the form of classifications and groupings). While SMT might translate particularly well the vocabulary items of a long sentence by using the context of directly neighboring words, it might translate not so well when it comes to overall grammaticality (e.g. long-distance grammatical problems, unseen morphological forms, etc.). Problems with GT misreading syntax, for instance, were noted by all the translators in Pym (2009), who "were generally appalled by the resulting wild mistranslations" (Pym 2009, p. 141).

To illustrate what can happen to the overall grammaticality of a statistical machine translation of a relatively syntactically complex sentence, Table 1 contains in the first row a single-sentence judgment summary taken from the Civil Division text. The second row contains an unedited Google translation of the judgment summary and the last row, a human-edited version we rendered by using as many correct words and phrases as possible from the Google translation. The errors pointed out result from the syntactic complexity of this single-sentence judgment summary.

Source text
6.- Derecho marítimo
6.1. La STS 28-9-2005 (RC 769/2005) destaca porque en ella, al examinar un supuesto de responsabilidad por abordaje, diferenciando sus distintas clases, se declara que, sin perjuicio de que las disposiciones contenidas en el Convenio de Bruselas de 23 de noviembre de 1910 sobre unificación de ciertas reglas en materia del abordaje, formen parte del ordenamiento jurídico español y sean de aplicación directa, resulta aplicable la legislación interna, con exclusión de cualquier otra, cuando los buques implicados son de nacionalidad española y el abordaje ha tenido lugar en aguas jurisdiccionales españolas.
Google Translate
6. - Maritime Law
6.1. The STS 28.09.2005 (RC 769/2005) stands out because in it, to consider a theory of liability for collision, differentiating their various classes, states that, without prejudice to the provisions of the Brussels Convention of 23 November 1910 on the unification of certain rules relating to the collision, part of the Spanish legal system and have a direct, domestic law applies to the exclusion of any other, when the vessels involved are of Spanish nationality and the approach has taken place in Spanish waters.
Human-edited version
6.- Maritime Law
6.1. The Judgment of the Supreme Court of 28-9-2005 (Appeal 769/2005) stands out because when it considers liability for collision by differentiating its various classes, it states that even though the provisions in the Brussels Convention of 23 November 1910 for the Unification of Certain Rules of Law with Respect to Collision Between Vessels are part of the Spanish legal system and are directly applicable, domestic law applies, to the exclusion of any other law when the vessels involved are of Spanish nationality and the collision has taken place in Spanish waters.

Table 1. Example of a Civil Division judgment summary accompanied by a Google Translate machine translation and a human-edited version. Highlighted passages indicate the source text grammatical and morphological areas that GT had problems with in the first row, the corresponding Google translations in the second row, and our human solutions and/or modifications in the human-edited version in the third row.

While the overall grammaticality of the GT output is mostly inadequate, most of the vocabulary items and the grammar of individual phrasal items are indeed correct, seeing how many such items could be left unchanged in our human-edited version. An interesting error, however, occurred with the lexically ambiguous word *abordaje*. It was translated correctly as 'collision' (the specialized meaning) the first time it appears and incorrectly as 'approach' (the everyday meaning) the second time it appears. Often, vocabulary in the everyday world takes on another meaning in the field of law. This vocabulary is often referred to as 'semi-technical', one of the areas of legal vocabulary we will describe in the following section.

Garcia (2010) is the only study that includes a text on legal topics[2]. GT might be acceptably accurate for other legal texts, although our text, unlike those in the other studies, is aimed at expert readers and much longer in length with 12, 263 words vs. an average of 199 words. Lee and Liao suggest for the future that "text genres of longer length may provide more in-

[2] The text used in Lee and Liao (2011) was a cellphone care instruction guide and the text in Pym (2009), a publication in an online dictionary of technical terms.

depth insights into MT use" (2011, p. 143). Our long text provides a sizeable sample of 621 legal vocabulary items on which this paper performs an in-depth empirical analysis of MT accuracy. To our knowledge, no such analysis has been performed yet.

3. Nature of the Vocabulary Items under Study

614 of the 621 vocabulary items under study can be classified as symbolic items, while the remaining 7 can be classified as functional items. Functional items are:

> grammatical words or phrases that have no direct referents either in reality or in the universe of concepts, but which serve to bind together and order those that do. Examples from the legal sphere are 'subject to', 'inasmuchas', 'hereinafter', 'whereas', 'concerning', 'under' and 'in view of'. Deictics, articles, auxiliaries, modals and other purely syntactic and morphological markers also belong with this group, as do other more complex units like 'unless otherwise stated', 'as in section 2 above', 'in accordance with order 14' and similar phrases (Alcaraz and Hughes, 2002, p. 16).

The 7 functional items are varyingly complex conjunctions or prepositional phrases such as *al régimen de* (under), *en atención a* (in view of), *según lo dispuesto en* (pursuant to), *ex artículo* (referred to in article). Certain functional items may have a high level of frequency in legal texts or be somewhat peculiar to them. Either way, the translator must be aware of them and know what their most contextually appropriate equivalents are in the target language. For example, 'pursuant to' is a frequent pattern among legal texts and, by extension, arguably characteristic of the discourse. In everyday language one would likely use 'according to' instead. In most cases, the individual units of multiword functional items cannot be translated one-to-one. The items are not compositional and equivalence must be established by the translator on the basis of probably all the co-occurrents in the phrase. For example, 'to the regime of', a literal word-for-word translation of *al régimen de*, would not make sense as a substitution for 'under' in 'under the contract', a possible translation of *al régimen del contrato*.

The 614 symbolic items "refer to things or ideas found in the world of reality, physical or mental" (Alcaraz and Hughes 2002: 16). 421 of them can be classified as terms and 193, as phrases. A term is defined as "a designation of a defined concept in a special language by a linguistic expression" and "can consist of single words or be composed of multiword strings. The distinguishing characteristic of a term is that it is assigned to a single concept, as opposed to a phrase, which combines more than one concept in a lexicalized fashion to express complex situations" (ttt.org-CSL Framework, 2001, ISO 12620 Data Categories). For example, *conocimiento de embarque* (bill of lading) is a term, whereas *bajo el régimen de conocimiento de embarque* (under a number of bills of lading) is a phrase.

The 614 symbolic items (421 terms, 193 phrases) can be divided into five vocabulary groups: purely technical vocabulary (221 terms, 110 phrases), semi-technical vocabulary (100 terms, 26 phrases), everyday vocabulary frequently found in legal texts (62 terms, 56 phrases), and official legal vocabulary (38 terms, 1 phrase).The purely technical, semi-technical, and official items can be classified under the following areas of law depending on the case: procedural law (148 cases), civil law (144 cases), commercial law (123 cases), constitutional law (18 cases), family law (12 cases), criminal law (10 cases), international law (7 cases), tax law (8 cases), European Union law (7 cases), administrative law (6 cases), inheritance law (5 cases), insurance law (5 cases), employment law (2 cases), and United Nations law (1 case).

In the following paragraphs each of the five types of vocabulary items are described in more detail.

Technical items, "whatever their origins may have been, now belong almost exclusively to the vocabulary of the law, or are firmly attached to this sphere in their everyday usage" (Alcaraz and Hughes, 2002, p. 154). Single-word examples (and their English translations) taken from the Civil Division text (and its English translation) include *demandado* (defendant), *exequátur* (exequatur), *homicidio* (manslaughter), *litigio* (litigation), *testamento* (will), *testador* (testator). While each of these items mostly has a single meaning, others "are more complex in that their precise senses may vary with context, though the overall field of reference remains the legal one" (ibid, p. 156). For example, *jurisprudencia* may be translated as 'jurisprudence' (if referring to the science of the law) or 'case law' (if referring to the decisions of judges relating to particular matters in contrast to statute law), and *embargo* as 'seizure' (in the context of civil law) or 'embargo' (in the context of international law). Technical multi-word terms and phrases "are meaningful only in a legal context, even though the individual words of which they are composed may belong to the general vocabulary of everyday speech" (ibid, p. 157). The following are multiword examples (and their English translations) taken from the Civil Division text (and its English translation): *cuestiones jurídicas suscitadas* (points of law that arise), *derechos reales* (rights in rem), *desestimación de la acción* (failure of the action), *desestimar el recurso interpuesto por* (dismiss the appeal brought by), *ejecución de una hipoteca* (foreclosure), *modificación jurisprudencial* (departure from precedent), *reparación íntegra del daño* (full compensation for damage), *resolución del contrato* (termination of contract). Some of the individual words of these multiword units may even have other meanings in other contexts. For example, other potential contextual meanings of *reales*, *reparación*, *recurso*, or *resolución* could be 'real', 'reparation', 'resource', or 'resolution', respectively. Some of these multiword items may only be properly understood by considering all the co-occurrents together and hence cannot be translated using a word-for-word approach. For example, *ejecución de una hipoteca* and *modificación jurisprudencial* would not make sense if rendered as 'execution of a mortgage' and 'jurisprudential modification', respectively.

Semi-technical vocabulary "is a more complex group, since it contains terms [and phrases] that have one meaning (or more than one) in the everyday world and another in the field of law" (ibid, p. 158-159). The items "belonging to this group are more difficult to recognize and assimilate than wholly technical terms [and phrases]" (ibid, p. 17). A phrase example (and its English translation) taken from the Civil Division text (and its English translation) is *causas de* (grounds for). Outside the sphere of procedural law it could be rendered as 'causes of'. Another phrase example is *beneficios derivados de* (profits arising from), which could be rendered as 'benefits deriving from' outside of the sphere of commercial law. A multiword term example is *masa activa* (insolvency estate), which could be translated as 'active mass' outside the domain of commercial law. Another multiword term is *acción infundada* (unmeritorious proceedings), which could be translated as 'unfounded action' outside the domain of procedural law. A few single-word term examples are *prenda* (pledge), *sociedades* (corporations), and *competencia* (jurisdiction), which in the everyday world could be translated as 'garment' (piece of clothing), 'societies', and 'competition', respectively.

Everyday vocabulary frequently found in legal texts consists of items "in general use that are regularly found in legal texts but, unlike the previous group, have neither lost their everyday meanings nor acquired others by contact with the specialist medium" (Alcaraz and Hughes, 2002, p. 18). Translators "will probably find that the terms [and phrases] are easier to understand than to translate, precisely because they tend to be contextually bound" (ibid: 162). A phrase example (and its English translation) taken from the Civil Division text (and its English translation) is *realidad y efectividad* (tangible effect). While each of the words in the phrase is being used in the most general sense, the literal rendering in English of each of them 'reality and effectiveness' might be awkward or nonsensical to the reader, and so it may

well be difficult for the translator to come up with an appropriately worded translation even though the source phrase is not difficult to comprehend. There are also everyday items that may have more than one everyday meaning depending on the context. A single-word term example is *omisión*, which in the Civil text document was translated as 'failing to act', not 'omission', which is how *omisión* would be translated were it being used according to its most common everyday meaning. The translator might not recall right away the less common meaning and might not be able to quickly produce a contextually appropriate translation. There are also everyday items that while they may not have a contextually sensitive meaning or wording, their register is high or their use is uncommon. A few examples of high-register items (and their English translations) taken from the Civil Division text (and its English translation) are: *homologar* (approve) instead of, say, *aprobar*; *decantarse por* (opt for) instead of, say, *favorecer*; and *coadyuvar* (contribute) instead of, say, *contribuir*. A few examples of not-commonly-used items (and their English translations) taken from the Civil Division text (and its English translation) are: *pleno* (plenary session), *su vigencia* (its period of operation), and *información veraz* (honest information).

Official legal vocabulary can include the names of specific laws, conventions, titles of legal professions or documents, etc. Official items that are part of a supranational entity often have official equivalents in other languages. The translator must already know or find out what they are. Some term examples (and their English translations) taken from the Civil Division text (and its English translation) are *Artículo 17 del Convenio de Bruselas de 27 de septiembre de 1968* (Article 17 of the Brussels Convention of 27 September 1968), *Convención de las Naciones Unidas sobre Contratos de Compraventa Internacional de Mercaderías, de 11 de abril de 1980* (United Nations Convention on Contracts for the International Sale of Goods of 11 April 1980), *Principios de Derecho Europeo de Contratos* (Principles of European Contract Law). While the official items that are part of the national legal system of Spain do not have official equivalents in English, there might be translations available in English that are time-tested and/or documented in important dictionaries or databases and thus often used by translators. Even so, English translations of the official vocabulary of the legal system of Spain do not have to be worded in any specific way as long as they are semantically correct and not "clumsy or too long to be practical" (Biel 2008, p. 34). Some examples (and their English translations) from the Civil Division text (and its English translation) are *Boletín Oficial de la Provincia* (Provincial Official Gazette), *Ley 27/1992, de Puertos del Estado y de la Marina Mercante* (Law No 27/1992 on National Ports and the Merchant Navy of 24 November 1992), *Texto Refundido de la Ley de Sociedades Anónimas* (consolidated version of the Law on public limited companies). Like the other categories, they may also include ambiguous words.

What particularly stands out is a high level of contextual determinacy either in how the source language items are meant to be understood or in how they should actually be written in the translation. This 'contextual sensitivity' seems a large part of what compels legal translators to spend time looking vocabulary items up in other resources, to verify how a source language item should be interpreted or how it should be written in the translation. The following brief paragraphs summarize what is specifically meant by 'contextually sensitive' as it pertains to each group of vocabulary items.

The largest group in which all the vocabulary items are contextually sensitive is that of semi-technical vocabulary (100 terms, 26 phrases). These items are grouped together because they have at least two different contextual meanings, a legal one and an everyday one, and, in theory, may be interpreted in more than one way.

The other group in which all the items are contextually sensitive is that of the 7 functional phrases, which may either need to be translated in a legally peculiar way or may only be properly understood on the basis of all the co-occurrents.

The 331 purely technical vocabulary items are contextually sensitive in 163 cases (98 terms, 65 phrases) either because single-word terms may have more than one legal sense or because multiword terms or phrases may contain ambiguous words or the multiword units themselves may only be properly understood on the basis of all the co-occurrents.

The 118 everyday items frequently found in legal texts are contextually sensitive in 51 cases (26 terms, 25 phrases). The category consists of single-word terms that may have more than one sense, multiword terms or phrases that contain ambiguous words or the multiword units themselves may only be properly understood on the basis of all the co-occurrents, as well as items whose translation needs to be worded in a specific manner. An example of a phrase that needed to be translated in a particular way is *comportamiento del adminstrador* (conduct of the administrator). *Comportamiento* in most cases is translated as 'behavior', but since the situation refers to one's behavior in a professional capacity, 'conduct' is preferred.

The 39 official vocabulary items are contextually sensitive in 23 cases (22 terms, 1 phrase) because either they have official equivalents in English that need to be worded in a specific manner (13 cases) or contain ambiguous words (7 cases).

On the basis of these parameters, context in 370 out of the 621 total cases (i.e., 60% of the cases) was assessed to be critical in how a source language item should be interpreted or in how its translation should be written.

4. Background: Documentation Experience

When in 2006-2007 we translated the Civil Division text, we prioritized using online EU multilingual databases containing English and Spanish language versions over any other non-EU resource when we were faced with translating the difficult vocabulary under study. Our main reasons for doing so were: 1.) to test how applicable EU resources could potentially be for translating the vocabulary found in a legal text produced by the Supreme Court of Spain (which is a member of the EU), 2.) because we wanted our English to be worded as close as possible to that used at the EU level, as the translation specifications did not include any particular English-speaking target culture and the translation was in theory being done for European readers in general (not necessarily native speakers of English), and 3.) for convenience purposes—EU resources are available free of cost and online.

595 (i.e. 96%) of the 621 vocabulary items were found to have a suitable equivalent in existing free online EU multilingual databases, such as EUR-Lex (Access to European Law) in 345 instances, IATE (inter-active terminology for Europe) in 247 instances, and the European Judicial Network (EJN) in 3 instances[3]. The other 26 cases could in 17 instances be documented in Alcaraz and Hughes's Spanish-English Diccionario de Términos Jurídicos (2005) and in the remaining other 9 instances, in various other non-EU resources.

In all cases we documented the longest piece of text possible. This was preferable because longer items include more context and consequently more information about how they

[3] EUR-Lex and IATE are the two multilingual resources that are used most by the translators of the largest translation organization in the world: the Directorate-General for Translation of the European Commission (European Commission, 2009, pp. 14-15). The multilingual corpus EUR-Lex, available online since 2002 (Liebwald, 2009, p. 274), provides free access to European Union law and other documents considered to be public. Released officially to the public in March 2007, the terminology database IATE "is a shared and interactive term base of the institutions and other bodies of the EU, designed as the main instrument for the multilingual drafting of Community texts" (Muñoz and Valdivieso, 2009, p. 375, translation ours). IATE and its beta version released in 2006 were available for use at the same time the translation of the Civil Division text was being done and during the editing phase. The EJN, for its part, contains information about Member States, their civil and commercial law, and European law.

need to be translated. As pointed out in the previous section, context in 60% of the cases was critical in how a source language item should be interpreted or in how its translation should be written. In any event, 463 out of the 621 cases (75%) contained more than one word (i.e., 200 phrases and 263 multiword terms). Terminological resources such as IATE, in the case of the EU resources, or Alcaraz and Hughes (2005), in the case of the non-EU resources, were consulted first, because they are more user-friendly. While a multilingual document repository like EUR-Lex might offer many possibilities (including terms and phrases that are not registered in a specific terminological resource), one can end up spending too much time sorting through too many parallel documents in its database. So it might be more efficient to first try using a high quality terminological resource, if one is indeed available (see Melby [2012] for an in-depth discussion of the role and importance of high quality term bases in the age of aligned multilingual corpora). Nevertheless, no matter how efficiently one tries to go about it, a good deal of time can be spent gathering, inputting, or sifting through written data.

5. Guarded Optimism

There are reasons to believe that SMT, "currently the dominant paradigm in MT research [with] a growing share of the MT market" (Forcada, 2010, p. 221), should be able to provide suitable translations for our sample of 621 researched terms and phrases. On the one hand, much of the currently available translated text used to train SMT comes from international organizations that often write about legal matters, such as the EU and the United Nations (Koehn, 2010, p. 53). On the other, SMT can use the immediate co-text of terms and phrases to identify translations that are likely correct for potentially ambiguous pieces of text whose meanings are hinged on the other words which surround them. SMT draws its solutions from an enormous database of previously translated texts where millions of sentences in one language have been aligned with their human translations in the other language. If provided the option of a longer phrase translation, SMT tends to use it (as we set out to do when documenting translation solutions back in 2006-2007), although longer phrases are less frequent and hence less statistically reliable (Koehn, 2010, p. 141). Even though a longer phrase translation is there, SMT might not select it, as there may not be enough statistics and hence enough reliable probability estimates.

In any event, the main hypothesis was that more often than not GT would be able to provide suitable translations for the 621 vocabulary items. There were also two sub-hypothesis: 1.) GT would in a good number of cases be able to provide suitable translations for 'contextually sensitive' pieces of text thanks to its phrase-based statistical approach. 2.) At the same time however, most of the errors would likely be produced when translating 'contextually sensitive' words or phrases. On the one hand, disambiguating ambiguous words and phrases has long been the Achilles' heel of MT ever since its first generation of systems. MT, however, has greatly improved thanks to the statistics of frequencies and corresponding probabilities drawing from a potentially comprehensive corpus of past translations and its ability to inform its decisions by using immediate co-text. On the other hand, there is no guarantee that the translation solutions provided by the open-domain GT will be worded in a desirable way. The database may not contain translations with a desired wording or the system may not heuristically be able to draw on them.

6. Methodology

While it would have been quicker for analysis purposes to feed only the vocabulary items under study into GT, the results in many cases would have likely not benefitted from access to

the surrounding co-text the system might be able to use as relevant input. Therefore, we fed the entire source text into GT, but one judgment summary section at a time for ease of data collection. A difference would not have been made in the MT results had the entire source text been fed into GT all at once. While SMT may on occasion be able to source from its database a translation for a previously seen sentence, its recognition of co-text does not stretch beyond the sentence, which is why it is said that MT translates a text "as a bag of unconnected sentences" (Hardmeier and Federico, 2010, p. 288).

Machine translations that were different from the handpicked ones were analyzed to determine whether they were correct. Though absolute synonymy in a given language is often thought not to be possible, translators may often be able to choose among several options to express the same thing in a legal translation (Mayoral, 2004, p. 61).

This study uses "a harsh correctness standard – is the translation perfect or not?" (Koehn, 2010, p. 218). There is a reasonable chance no MT mistakes are made at the term and phrase level. Because "correctness may be too broad a measure", the study breaks it down into "the two criteria fluency and adequacy"; the former "involves both grammatical correctness and idiomatic word choices", while the latter asks: "Does the output convey the same meaning as the input […]? Is part [or all] of the message lost, added, or distorted?" (Koehn, 2010, p. 218). Correct renditions are judged to be both 'fluent' and 'adequate', while incorrect ones are judged to be either 'influent' or 'inadequate', or both. While a harsh standard may overlook 'incorrect' translations of multiword items that are not entirely incorrect, it discourages the leniency variability that can occur with multiple-range scoring. Moreover, partially incorrect output would in theory not spare one from having to do research, though it might prompt one to devise a more efficient research strategy.

When an item being used in the same way more than once was machine translated differently on more than one occasion, the MT variations were recorded, although the item was classified according to the 'best' possible variant. For example, if a particular item was translated incorrectly in one instance but correctly in another, it was counted as correct under either the category of the 'same' translations or the 'different but correct' ones, depending on the case. If a machine translation of an item was 'different but correct' in one instance and the 'same' in another, that which was the same was counted[4]. Different source-text contexts and the use of free online SMT on different occasions can prompt different machine translations. When an item appearing more than once was consistently machine translated in the same way, this was not taken stock of[5].

7. Results

223 of all the results (36%) were machine translated 'in the same way', 177 (28.5%) were machine translated 'in a different but correct way', and 221 (35.5%) were machine translated 'incorrectly'. That a little over 64% of the results were machine translated correctly supports our main hypothesis that more often than not GT is able to provide suitable translations for the 621 items. In Figure 1 the percentages of these overall results can be compared with those of the results of each of the 5 vocabulary categories under study.

[4] Opting for the machine translation that was the same as our previously documented translation solution helped speed up data collection, as ascertaining whether a different machine translation was correct would often entail more investigative work on the part of the researcher. In other words, our handpicked solutions were not 'better' by default than any other possible correct MT rendition; they were merely less burdensome to categorize.

[5] This measure was also less burdensome for data collection purposes.

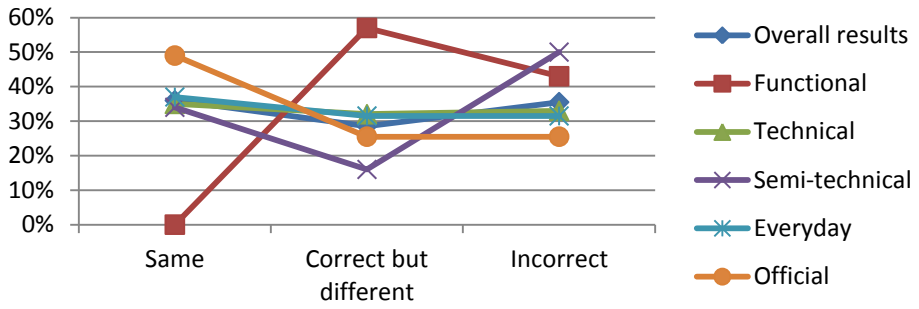


Figure 1. Percentages of the overall results and those of the results of each of the 5 vocabulary categories under study.

The technical and everyday items are the two categories whose results are most consistent with the overall ones. They are significant in volume; the former category, the largest of all the other categories with its 331 items, comprises some 53% of the entire sample under study and the latter category with its 118 items, some 19%. The results of the remaining three categories deviate most from the overall ones. The functional and the official items, due to their limited volume: 7 (1%) and 39 (6%), respectively, might not be the most statistically reliable and quite possibly random. However, the semi-technical items are reasonably significant in volume, with 126 units making up some 20% of the entire sample. Either way, both the semi-technical and the functional items are the two entirely 'contextually sensitive' categories and they have the highest error percentages. It is also worth pointing out that in each of the remaining categories (i.e., those in which not all the items are contextually sensitive) the 'incorrect' results involved the highest concentrations of contextually sensitive vocabulary. Figure 2 compares the percentages of contextually sensitive items in each area of the results of the remaining vocabulary categories with the overall results.

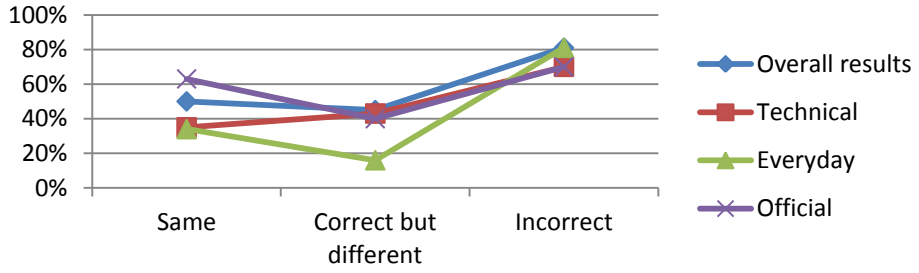


Figure 2. Percentages of 'contextually sensitive' items in each area of the overall results and the results of the remaining vocabulary categories.

The overall results in Figure 2 show that in a little over 80% of all the cases in which GT translated incorrectly, contextually sensitive items were involved. This absolutely supports our second sub-hypothesis that most of the errors would likely occur when translating contextually sensitive items. But looking at the 370 total 'contextually sensitive' items as a whole (i.e. 60% of the entire vocabulary sample) reveals that 111 of them were machine translated 'in the same way', 80 'in a different but correct way', and 179 'incorrectly'. That a little over half of them were machine translated correctly supports our first sub-hypothesis that GT would in a good number of cases be able to provide correct translations for 'contextually sensitive' items. Accuracy in as many as around half the cases was not expected.

Take as an example *según lo dispuesto en*, a functional item originally documented in EUR-Lex as 'pursuant to'. GT translated it correctly as 'under the provisions of'; it remarkably did not grind out a word-for-word error to the likes of 'according to that available in'. Another example is *propietario y explotador*, a contextually sensitive technical phrase from commercial law that was originally documented in EUR-Lex as 'owner and operator' and then machine translated in the same way. In other contexts *explotador* could be translated as 'exploitative' or 'exploiter', for example. It seems GT was able to use *propietario* as a contextual cue. A couple of error examples, however, are the semi-technical procedural law terms *proceso de ejecución* and *acción ejecutiva*. The first was originally documented in EUR-Lex and the second, in Alcaraz and Hughes (2005). The translation solution arrived at in each case was 'enforcement proceedings', as it turns out the terms were being used synonymously. GT mistakenly rendered them according to their non-technical meanings as 'implementation process' and 'executive action', respectively. More contextually sensitive examples are described in the Appendix.

8. Conclusion

The results of this study show that GT could indeed translate accurately vocabulary taken from a voluminous legal text aimed at expert readers in a little over 64% of the cases. Little time is invested in using this free online SMT system. It produces translations for words and phrases instantly. That it is able to do so correctly at least more often than not in an area of translation widely considered one of the most difficult renders it, in our opinion, a considerably reliable tool legal translators might indeed find beneficial. I've yet to come across a tool that is (close to) 100% effective, let alone in the legal translation context. SMT may come in especially handy when it is able to correctly translate the difficult vocabulary one might otherwise need to spend time looking up. Manually looking up terms and phrases over and over again can take up a good deal of time and ideally should be kept to a minimum. A reliable SMT system might significantly cut down on the amount of vocabulary one might need to look up.

When we translated without MT in 2006-2007, MT was, to put it mildly, not thought of as a viable source of translation support, let alone for legal texts, often characterized by contextually sensitive vocabulary and complex discourse structures. Context in a little over 80% of the MT errors reported in this paper was indeed a factor, although a little over half of all the 'contextually sensitive' cases were machine translated correctly. According to these results, MT's contextual Achilles' heel still remains vulnerable, but certainly not entirely at a little less than 50% thanks to recent phrase-based statistical methods. In any event, one must keep in mind that as long as the database contains enough translations that are similar to the text being translated, the less ambiguous the co-text of the input and the likelier it is to be a frozen pattern of language, the likelier it is SMT will translate accurately in terms of correct word-sense disambiguation. The co-text that can be read by the machine will likely be longer and hence more likely to be translated according to its contextually determined meaning.

It is hoped that this paper will shed light on the quality of free online SMT in the legal translation context. Having an idea of the aspects in which it might perform particularly well or bad in a specific context is important so that one may effectively work with the output if one so chooses and set reasonable expectations for it. This paper demonstrates with a sizeable sample of legal vocabulary, considered "the greatest single difficulty encountered initially by legal translators" (Alcaraz and Hughes, 2002, p. 16), that free online SMT might perform consistently well in this area. Legal translators, especially those that are new to the field, might find that SMT definitely has something to offer in terms of vocabulary.

References

Alcaraz, E. and Hughes, B. (2002). *Legal Translation Explained*. St. Jerome Publishing, Manchester.

Alcaraz, E and Hughes, B. (2005). *Diccionario de términos jurídicos*, 8th ed. Editorial Ariel, Barcelona.

Biel, L. (2008). Legal Terminology in Translation Practice: Dictionaries, Googling or Discussion Forums? *Skase*, 3(1):22-38.

Cao, D. (2007). *Translating Law*. Multilingual Matters Ltd, New York.

Consejo General del Poder Judicial. (nd). Crónica de jurisprudencia. http://www.poderjudicial.es/cgpj/es/Poder_Judicial/Tribunal_Supremo/Actividad_del_TS

Forcada, M. L. (2010). Machine Translation Today. In Gambier, Y. and Doorslaer, L.V. (eds) *Handbook of Translation Studies Volume 1*, pages 215-223, John Benjamins, Amsterdam.

Garcia, I. (2010). Is Machine Translation Ready Yet?. *Target*, 22(1):7-21.

Hardmeier, C. and Marcello, F. (2010). Modelling Pronominal Anaphora in Statistical Machine Translation. In *Proceedings from the 7th International Workshop on Spoken Language Translation*, pages 283-289, Paris.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York.

Lee, J. and Liao, P. (2011). A Comparative Study of Human Translation and Machine Translation with Post-Editing. *Compilation and Translation Review*, 4(2):105-149.

Liebwald, D. (2009). Interfacing between Different Legal Systems Using the Examples of N-Lex and EUR-Lex. In Grewendorf, G. and Rathert, M. (eds) *Formal Linguistics and Law*, pages 257-291, Mouton de Gruyter, Berlin.

Mayoral, R. (2004). Lenguajes de especialidad y traducción especializada: la traducción jurídica. In Gonzalo, C. and García, V. (eds) *Manual de documentación y terminología para la traducción especializada*, pages 49-71, Arco/Libros, Madrid.

Melby, A. (2012). Terminology in the Age of Multilingual Corpora. *JOSTRANS*, 18:7-29.

Monzó, E. (2008). Documentación para la traducción inglés-español. In Ortega, E. (ed) *La traducción e interpretación jurídicas en la Unión Europea*, pages 733-752, Comares, Granada.

Muñoz, M. and Valdivieso, M. (2009). La terminología en las instituciones de la UE: de la fragmentación a la convergencia. *EntreCulturas*, 1:365-383.

Pym, A. (2009). Using Process Studies in Translator Training: Self-Discovery through Lousy Experiments. In Göpferich, S., Alves, F., and Mees, I.M. (eds) *Methodology, Technology and Innovation in Translation Process Research*, pages 135-156, Samfundslitteratur, Copenhagen.

Pym, A. (2013). Translation Skill-Sets in a Machine-Translation Age. *Meta*, 58(3):487-503.

ttt.org-CSL Framework. (2001). ISO 12620 Data Categories. http://www.ttt.org/clsframe/

Appendix. GT Machine Translations of Contextually Sensitive Vocabulary

Rendered in the Same Way

A semi-technical example is *balance de situación*, a commercial law term that was originally documented in IATE as 'balance sheet' and then machine translated in the same way. It is remarkable that it was not machine translated literally and incorrectly as 'balance of situation/situation balance'.

Omisión del deber de is an everyday phrase that was originally documented in EUR-Lex as 'failure to'and then machine translated accordingly. An awkward or potentially nonsensical word-for-word translation might have been 'omission of the duty of'. Another everyday example is *vicio oculto*, a term originally documented in IATE as 'latent defect' and then machine translate in the same way. It is noteworthy that GT did not translate the two-word string as 'hidden vice', for example, 'hidden' and 'vice' being appropriate translations of *oculto* and *vicio*, respectively, in most other contexts.

An official term example is the European Union *espacio económico europeo*. Its official equivalent in English: 'European economic area', which we documented in IATE, was the MT of GT. Obviously, GT was able to use the co-text of this multiword term and match it to relevant examples existing in its database. A word-for-word incorrect rendition might be "European economic space" or "European financial space", as "space" and "financial" are often translations of *espacio* and *económico*, respectively.

Rendered in a Different but Correct Way

A technical term example is *abuso de derecho*. It belongs to procedural law and was originally documented in IATE as 'misuse of law' and then machine translated differently but correctly as 'abuse of process'. In most contexts *derecho* may be translated as 'right' or 'law'. Nevertheless, GT was able to decide on the suitable translation of *derecho* by using the relevant co-text: *abuso de*. Another technical example of a procedural law item containing *derecho* is the phrase *ajustado a derecho*, which was originally documented in EUR-Lex as 'legally sound' and then machine translated as 'consistent with the law'. Had GT rendered *ajustado a* as 'adjusted to', which would be correct in the majority of other contexts, the MT would be unidiomatic and hard to understand.

An everyday term whose tranlsation has to be carefully worded is *el elemento determinante*. We produced 'the decisive factor' thanks to EUR-Lex. GT, though, came up with 'the determinant'. Other suitable renditions might be 'the crucial element' or 'the determining factor'. But 'the determining element', an easily comprehensible word-for-word rendition, would be unidiomatic.

Rendered Incorrectly

An example of an everyday term is *situaciones consolidadas*, which we documented in EUR-Lex as 'previous situations'. GT's word-for-word rendition 'consolidated situations' is too difficult to make sense of. Another everyday term is the lexically ambiguous *fórmula*, which appears in the phrase *Especifica la Sala que no le da a dicho requisito la consideración de fórmula expresa y solemne*…. We managed to document *fórmula* in Alcaraz and Hughes (2005) as 'wording', as in 'The Chamber specifies that it does not give consideration to that requirement in express and solemn wording…'. GT, however, made the mistake of translating it as 'formula', which in most other contexts would be appropriate.

Texto Refundido de la Ley de Sociedades Anónimas is an official commercial law term we documented in EUR-Lex as 'Consolidated Version of the Law on Public Limited Companies'. GT, on the other hand, came up with 'Consolidated Companies Law', which appears in a good number of sentence examples crawled by Linguee (which can be thought of as a sort of 'Google of bilingual text'). A potential problem might be that its inclusion of 'Companies' and not 'Public Limited Companies' is unnecessarily too general a solution, whereas 'company' is normally the equivalent of *sociedad* when it is unmodified by an adjective and 'public limited company', normally that of *sociedad anónima* in the European context (e.g. in the United States one might prefer 'joint-stock company'). Nevertheless, as the EUR-Lex database continues to grow it is now possible to find more English translations of the source term that employ 'Companies' instead of 'Public Limited Companies' (e.g. 'Amended Text of the Law on Companies'). So perhaps our categorization is nitpicky or our hand-picked solution runs the risk of being clumsy or too long to be practical.

Towards desktop-based CAT tool instrumentation

John Moran moranj3@cs.tcd.ie
Christian Saam saamc@cs.tcd.ie
Dave Lewis dave.lewis@cs.tcd.ie
Department of Computer Science and Statistics, Trinity College Dublin, Ireland

Abstract

Though a number of web-based CAT tools have emerged over recent years, to date the most common form of CAT tool used by translators remains the desktop-based CAT tool. However, currently none of the most commonly used desktop-based CAT tools provide a means of measuring translation speed at a segment level. This metric is important, as previous work on MT productivity testing has shown that edit distance can be a misleading measure of MT post-editing effort. In this paper we present iOmegaT, an instrumented version of a popular desktop-based open-source CAT tool called OmegaT. We survey a number of similar applications and outline some of the weaknesses of web-based CAT tools for experienced professional translators. On the basis of a two productivity test carried out using iOmegaT we show why it is important to be able to identify fast good post-editors to maximize MT utility and how this is problematic using only edit-distance measures. Finally, we argue how and why instrumentation could be added to more commonly used desktop-based CAT tools that are paid for by freelance translators if their privacy is respected.

1. Introduction

To measure the utility of machine translation (MT) in a post-editing context two dimensions must be measured - translation quality and speed. In large commercial post-editing (PE) projects weighted scorecards methods for quality control like the LISA QA Model and SAE J2450[1] can be adapted to assess the quality of post-edited machine translation (MT) without changing the underlying software (often spreadsheets). In particular to renegotiate word prices for post-editing it is important to know how fast a translator works in an MT post-editing context and when translating from scratch as human translation (HT). Unfortunately, measuring speed presents a challenge. It is possible to measure and validate translation and post-editing speed for in-house translators where conditions can be controlled. However, 74% of translators in Europe are freelance translators (Pym, Grin, Sfreddo, & Chan, 2013) and it is likely the percentage is high in other territories also.

The majority of these translators use sophisticated desktop-based Computer-Aided Translation (CAT) tools they have purchased themselves like Trados[2], MemoQ[3] and Wordfast[4]. However, none of these CAT tools record translation speed (Aziz et al., 2012). As a result, a common practice in translation agencies is to use edit distance between raw and post-

[1] http://www.sae.org/standardsdev/j2450p1.htm
[2] www.sdl.com
[3] www.kilgray.com
[4] www.wordfast.net

edited MT[5,6] to infer the utility of MT. In a paper that describes how post-editing was introduced in a Spanish translation agency over a period of years Silva (2014) neatly summarises this problem:

> *"One of the problems faced when introducing MT was that measuring its benefits on productivity on a representative number of translators over long periods of time was very difficult to achieve. Almost all translators employed by the company were freelancers (97%), mostly working remotely from home (88%)….Although online CAT tools are becoming more popular which may allow for easier productivity rate and post-editing effort tracking, most translators work on a locally installed CAT tool that does not offer such functionalities. The only information available was their feedback. As valuable as it is, it did not really provide valid post-editing effort figures. Internal translator productivity, on the other hand, was easily measured as there was complete control over their working environment and number of hours employed."*

In this paper we will survey a number of web-based and desktop-based systems that can be used to measure translation speed. We will then discuss some of the weakness of current web-based CAT tools from a translator perspective and how OmegaT[7], a free open-source desktop-based CAT tool that is used by many translators worldwide addresses some of these weaknesses. We will then present iOmegaT, an adapted version of OmegaT to which we have added instrumentation to log User Activity Data in a format we call CAT-UAD and describe the workflows for the system. We describe how MT utility can be measured by means of Segment Level A/B (SLAB) testing, in which CAT-UAD is recorded unobtrusively, and hence cheaply, as a translator processes segments in (A) HT segments and (B) MT segments. We will illustrate how data gathered by this means can be useful for translator selection for post-editing (PE) and discuss some other uses of temporal PE data. Finally, we will discuss some of the issues that may need to be addressed if instrumentation is to be adopted by other more commonly used desktop-based CAT tools normally purchased by translators.

Related Work

In general we will categorise related work on measuring translation and post-editing speed into applications that are obtrusive to the task of translation and those that are unobtrusive. We define an obtrusive translation environment as one in which a translator is asked to work in manner that is not business-as-usual. For example, in an obtrusive environment a translator may be asked to carry out an annotation task to score MT in terms of adequacy and fluency, the environment may use a constrained method for segment navigation or it may not be possible to leverage translation memory. Though some obtrusive tasks like annotation are useful, obtrusive environments reduce the quantity of translation speed data that can be gathered in a typical commercial translation scenario due to cost considerations (Lewis & Moran, 2010).

1.1. Obtrusive applications that can measure translation speed

The first use of UAD to record how translators interact with an editing environment can be traced back to TransLog (Jakobsen & Schou, 1999). In this application and its successor TransLogII (Carl, 2012) UAD is recorded in XML for subsequent analysis and replay. TransLog is designed for lab use. It is impractical for large-scale or multi-day/multi-language MT

[5] http://www.yamagata-europe.com/en-gb/blog/item/909/quality-metrics-for-machine-translation-output
[6] http://www.alphacrc.com/news/newsitem_29-11.php
[7] www.omegat.org

productivity testing. This lab or experimental focus makes it particularly suitable for use with eye-tracking software, which is used to examine cognitive aspects of the translation process.
Another desktop-based free and open-source application used to carry out translation research is PET (Aziz et al., 2012). The focus here is on MT post-editing. Translators can annotate MT proposals, e.g. in terms of quality using Likert scores. It also records translation speed and reports on edit distance. Another, more standards-based free open-source system that can be used to annotate post-edited MT is Ocelot[8] and two web-based MT rating systems that can also report on PE time are TransCenter (Denkowski & Lavie, 2012) and ACCEPT[9].

A number of web-applications to measure post-editing speed in a SLAB testing scenario have been described in the post-editing literature. CrossLang is a company that provides MT consulting and software services. They have developed a web application that can be used by their clients to measure post-editing speed and compare it to HT speed (Depraetere, De Sutter, & Tezcan, 2014). Autodesk also describe a similar internal system (Plitt & Masselot, 2010). A similar commercially available system that can be used to measure translation speed (and annotate segments) is the TAUS DQF platform[10]. This application is available for a monthly fee.

Features these three applications have in common are that translators can press a Pause button when they chose to take a break. In the Autodesk and TAUS DQF applications translators navigate through a text by means of a Previous and Next button. Thus, for example, if a translator is 200 sentences into a translation and decides to return to the 50th sentence to make a change based on new context, he must press the Previous button 149 times. There is no mention of self-review in publications that use this design so presumably translators carrying out productivity tests navigate the text from start to finish in a single pass. In the Crosslang application only a Next button is available so self-review is technically impossible. This kind of linked-list navigation can impact negatively on translation consistency, as the translator is not working in a business-as-usual manner in which they are free to navigate freely between segments. Also, no application listed above allows the translator to leverage translation memory. As such they do not work well in a typical translation agency workflow.

1.2. Unobtrusive applications that can measure translation speed

CASMACAT[11] and MateCAT[12] are two cooperating web-based CAT tool projects still in progress. While CASMACAT provides a more constrained working environment and is focused on research use, MateCAT is designed to be used by working translators, albeit "*volunteer translators and professional translators without advanced technical skills*" (Alabau et al., 2013).

Finally, IBM TM/2 is an offline CAT tool used internally in IBM[13]. A number of years ago a feature called MTLog was added to the tool to measure translation speed at a segment level. It is both desktop-based and also has a logging feature that has been added to an existing CAT tool. While results from this system have been presented orally[14] no published re-

[8] http://open.vistatec.com/ocelot
[9] http://www.accept.unige.ch
[10] https://www.taus.net
[11] www.casmacat.eu

[12] www.matecat.com

[13] www.ibm.com
[14] www.localizationworld.com/lwbar2011/presentations/files/E6.ppt

search is available on the topic. To the best of our knowledge, the system is only available for use internally within IBM and to its suppliers. Unfortunately, the MTLog feature is not available in the open-source version of TM/2 called Open TM2[15]. Table 1 presents a summary of this survey.

Name	Obtrusive	Architecture	Internal to company
Translog I&II	Yes	Desktop	No
PET	Yes	Desktop	No
TransCenter	Yes	Web	No
Autodesk tool	Yes	Web	Yes
Crosslang	Yes	Web	No
TAUS DQF	Yes	Web	No
IBM TM/2+MTLog	No	Desktop	Yes
CasmaCAT/ Mate-CAT	No	Web	No
ACCEPT	Yes	Web	No
iOmegaT Workbench	No	Desktop	No

Table 1. Survey of applications that can measure translation speed

As iOmegaT is a logging feature added to an existing desktop-based CAT tool functionally it is most like IBM TM/2 + MTLog. However, this system is not available outside of IBM. In terms of goals and functionality MateCAT is currently the most similar *available* application to iOmegaT. For this reason we will now examine some of the differences between web-based and desktop-based CAT tools as they appeal to quite different translator demographics.

2. Browser-based CAT tools – a discussion

In this section of the paper we take a step back from MT post-editing productivity testing and examine the purported trend towards web-based CAT tools in the previous quote by Silva. In a paper entitled *Power-shifts in web-based translation memory* Garcia (2008) notes that while online or web-based workflow systems may serve other stakeholders in the translation supply chain, they can involve drawbacks for freelancer translators. Amongst other issues he argues that web-based translation memory tools do not allow translators to access their own private locally stored translation memories and terminology databases. Access to locally stored terminology files does not just improve translation quality by ensuring consistency and reusing previous terminology research. Much anecdotal evidence suggests auto-complete using well-populated termbases can also aid translation speed.

Though this paper was published several years ago, we are aware of no browser-based CAT tool that facilitates access to local files. This is unfortunate as the Google Chrome browser allows access to local files[16] - so from a programming perspective it should be possible. Access to the local file system by the web-application seems necessary as we feel at least some translators would be unwilling to upload private termbases[17] to any third party. Aside

[15] www.opentm2.org

[16] https://developer.chrome.com/apps/fileSystem

[17] Private translation memory is less of a concern as a translator can always have a second desktop-based CAT tool open for concordance searches and full-sentence matches across clients are rare.

from the competitive advantage private termbases represent in terms of quality and translation speed, the contents of these databases may originate from various clients. Uploading them to any third party could be a breach of client confidentiality as defined in non-disclosure agreements commonly signed by translators.

Garcia also points to the fact that most browser-based systems do not let the translator retain bilingual assets generated while translating in the CAT tool. This means previous work cannot be referenced. In his summary he refers to web-based CAT tools as a "*looming reality"* that will "*alienate the most able section of its workforce and discourage promising newcomers from entering"*.

However, Garcia's pessimism may have been premature. According to recent surveys of CAT tools used by professional translators[18,19], by far the dominant architecture remains the desktop-based CAT tool. The 8 most used CAT tools in Figure 1 are desktop-based CAT tools. 3000 translators responded to the survey on proz.com, a translator website. Only 4 of the 16 systems listed feature a browser-based CAT tool. These are XTM[20], MemSource[21,] Wordbee[22] and Google Translator's Toolkit[23]. Interestingly, though they started out as applications that include a web-based CAT tool, XTM and MemSource now provide their own desktop-based CAT tools. Only Google Translators Toolkit, a free online service provided by Google, does not provide an export facility for linguistic assets (XLIFF/TMX/TBX files) or TIPP[24] export facility. Unfortunately, this export means the translator can no longer collaborate using a shared translation memory in parallel with other translators so the benefits of being online are lost.

A concern not addressed by Garcia is that translators who are used to working with a CAT tool for many clients may suffer a loss of productivity if they are asked to work in a new unfamiliar environment intermittently.

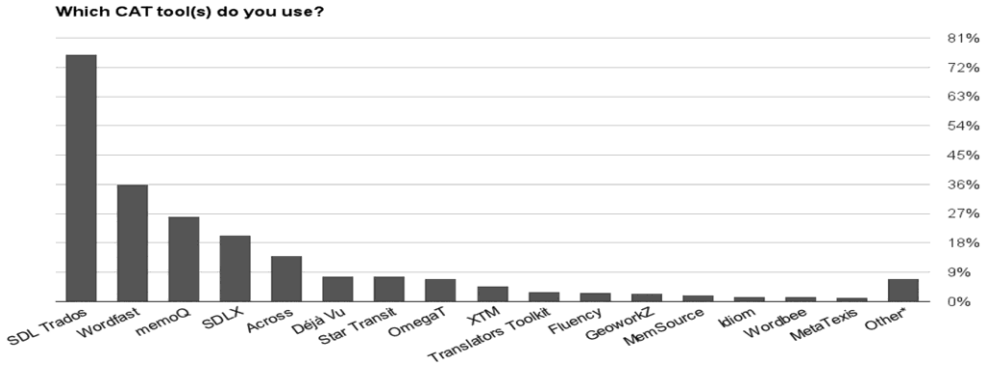


Figure 1. Results of a survey on CAT tool use (reproduced with permission).

[18] http://www.translationtribulations.com/2014/01/the-2013-translation-environment-tools.html
[19] http://prozcomblog.com/2013/03/28/cat-tool-use-by-translators-what-are-they-using
[20] http://xtm-intl.com
[21] http://www.memsource.com

[22] http://www.wordbee.com
[23] https://translate.google.com/toolkit
[24] https://code.google.com/p/interoperability-now

Finally, a shortcoming of the web-based tools surveyed above, including MateCAT is that translators cannot work if the connection to the Internet is lost. iOmegaT was developed in close collaboration with a large translation agency called Welocalize[25]. Feedback from translators working for the agency identified this as a frustration with existing web-based CAT tools provided by other agencies.

In short, web-based CAT tools provided for free by agencies or translation buyers and used intermittently by translators can be costly to those translators in terms of lost productivity or translation quality.

2.1. OmegaT

OmegaT addresses most of these concerns. It is a well-featured desktop-based CAT tool already used by many experienced professional translators. It can be used for offline and collaboratively using the team function and translators can use their own translation memories and terminology databases safe in the knowledge that the contents will not be uploaded to a third party. A feature of OmegaT's GPLv3 license is that any user of the software has the right to the source code for the software. This means assurance can be sought in this regard by reading the code or consulting with someone else who can. This is not possible for proprietary software where oftentimes complex legal agreements must be taken on face value. However, integration of instrumentation does not solve the problem of productivity loss due to a lack of familiarity for users of CAT tools. We will return to this problem in Section 5.

3. The iOmegaT Translator Productivity Workbench

The iOmegaT Translator Productivity Workbench is a fully-offline suite of software applications used to record and analyze how translators interact with a CAT tool, in particular when post-editing MT. Unlike web-based environments its offline nature means that test results cannot be affected by local network conditions or high server load. Static logging has no impact on the performance of OmegaT. The suite of applications is composed of four main components shown in Figure 2. These are:

1) The instrumented CAT tool (iOmegaT)
2) Middleware, used to transfer files to and from SDL TMS and SDL WorldServer
3) A collection of utilities to prepare a productivity test.
4) The analysis application and prototype replayer component.

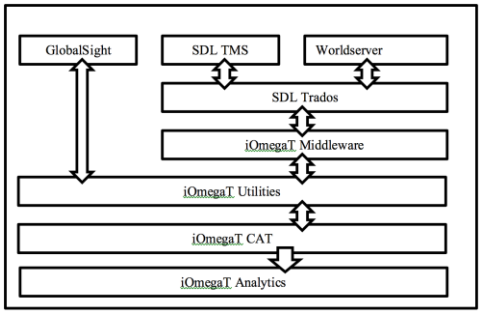


Figure 2. The iOmegaT Translator Productivity Workbench workflow

[25] www.welocalize.com

The software is designed to work in a typical large-scale or enterprise post-editing project workflow where there is some degree of content homogeneity within work packages that are handed off to translators or across multiple work packages on a single account. The workflow is outlined in Figure 2. In a typical scenario an engineer responsible for MT or a project manager selects a project from a TMS to use for productivity testing. If the TMS is GlobalSight only one step is required to prepare the project for productivity testing, namely deleting segments at random from the TMX file used to store the machine translation. This creates the HT segments for the HT/MT SLAB tests. If the files originate from an SDL TMS or SDL WorldServer system we use the Middleware applications to import the various dialects of XLIFF into OmegaT via Trados (sdlxliff) before deleting the MT segments using the iOmegaT Utilities. Initially, MT productivity tests were one way. Files could not be returned to the TMS they came from so they could not be used. This massively reduces the cost of productivity tests as prior to this translations were discarded. We expect the cost is of acquiring PE speed data to fall further where OmegaT is used as a production CAT tool. This means post-editing productivity can be measured on an ongoing basis rather than in a shunted test workflow.

The code we have added to OmegaT writes a stream of events corresponding to user activity and application context to XML log files stored on a translators PC as the translator works in the CAT tool. We call these XML log files instrumentation files to distinguish them from application log files. As opposed to application log files, instrumentation files are repayable in the Replayer. Currently, this component is an early prototype. It can only reply CAT-UAD at a segment level. Each time a source file is opened a new instrumentation file is created. CAT-UAD is mainly expressed as events. These events contain timestamps in milliseconds and contextual information. For example, in the case of a keystroke event the time of the event is recorded along with the key in question and the cursor position in the segment. Segment editing sessions are used to distinguish between the first and subsequent times a translator visits a particular segment. Figure 3 shows an example of a segment in which an MT segment is post-editing in 42 seconds. The project name, file name, sourceIndex and translatorID combine to uniquely reference a Source segment and this segment may be accessed or visited many times. In productivity tests we find on average translators visit each segment between two and three times.

```
<SegmentSession sourceIndex="14">
 <SourceText>This section describes the options you can use with the <x0/>cmdjob <x1/>command.</SourceText>
 <PreeditText> </PreeditText>
  <Events>
   <LogEvent Action="segmentOpen" Time="1363874598949"/>
   <LogEvent Action="mtMatchPlacement" In diesem Abschnitt werden die Optionen beschrieben, können Sie mit dem Befehl cmdjob."
   MTsystem="123456789" Time="1363874599029"/>.
   <LogEvent Action="segmentClose" Time="1363874641632"/>
   </Events>
   <PostEditTarget>In diesem Abschnitt werden die Optionen beschrieben, die Sie mit dem Befehl cmdjob verwenden können.</PostEditTarget>
<Comment postEditTime="42"/>
</SegmentSession>
```

Figure 3. A shortened example of a 42 second segment editing session

Once the translation is ready the translator zips and return that package by e-mail or FTP. The by the MT engineer or project manager runs the analysis software across the instrumentation files for all returned packages in a single batch operation. This populates a database that contains data on each editing session. On the basis of that database a summary of results for the current test is output as a spreadsheet. The report contains information on the number of words in each segment and the number of words that were discarded due to inactivity on the translator's part. Some translations are also discarded for other reasons, e.g. if a translator saves an empty target segment to the project translation memory. Cut-offs due to inactivity are defined in the analytics configuration file. For example, a cut-off of 5 minutes was used

for the productivity tests discussed below. For patent translation with much longer sentences 15 minutes is used. At an early stage in the design we decided against a Pause button, as it is not possible to verify if it was used and it conflicts with the unobtrusive philosophy underlying the design of the software.

4. Some results from two large scale productivity tests

Economically, from a translator's perspective temporal post-editing effort is the more important measure of effort (Krings, 2001). In this section we will discuss the results of two MT productivity tests carried out using iOmegaT, with a focus on temporal data. We will show that reliance on edit-distance alone is not advisable when measuring the utility of MT or selecting translators to work on post-editing projects.

	Dell	**Autodesk**
Total Number of Segments	14686	13145
HT	4920	1672
100% match	412	20
Already translated	2855	2249
Fuzzy Match	2854	2070
MT Changed Segments	3353	5703
MT Unchanged Segments	292	1431
MT Changed WPH	476	560
MT Unchanged WPH	942	837
MT Utility	+5%	+54%[26]

Table 2. Summary of data from two similar productivity tests

Table 2 shows a summary of statistics from two similar productivity tests carried out using iOmegaT for Autodesk and Dell, two companies that already use MT on a large scale. Translators were all experienced professional translators and MT had been in use for some time on the accounts in question so translators had passed many successive QA cycles. Spot-checking was applied to ensure basic quality levels were maintained in the productivity tests. In general segments were sentences. In both projects the mean sentence length was nine words and the mean number of visits per segment was just under two. Whether a segment was MT or not had no meaningful impact on the mean visit count.

100% match segments were segments in the translation memory provided to the translators before the project was started. "Already translated" segments are those the translator has translated or post-edited during the project and which appear again as full matches. "Fuzzy Match" segments were fuzzy or partial matches from the translation memory provided to the translator or the project translation memory that is populated as the translator works. "MT Changed" segments are segments that are changed by the translator and "MT Unchanged" segments were judged to require no changes. "MT Utility" is the overall utility of the MT in terms of speed relative to HT translation speed. Unsurprisingly a high MT utility corresponds to a high number of unchanged MT segments as unchanged MT segments required much less time to process. This can be seen in the Words per Hour (WPH) values for each category. All WPH values also account for self-review time across multiple segment visits or sessions. This is discussed in more Moran, Lewis, & Saam (2014).

[26] French, Italian, German, Spanish only

A feature of these two early productivity tests was that the two-day translation projects were handed off to two translators per language to be redundantly translated. However, they were free to translate normally using OmegaT, so although all translators started at the start of the job there was no stipulation that they should translate the same segments and some translators translated non-contiguously. Also, due to multithreading in OmegaT a segment may be presented to translator A as MT while the same segment may be shown to translator B as a translation memory match. This was recorded correctly by the instrumentation but was only apparent upon analysis. For some visualisations of segment progression and segment category distribution see Moran et al. (2014). In general faster translators translated more segments over the two days. Figures 3 shows PE speed compared against translators for both tests. There was a low positive correlation between translators in terms of PE speed. The Pearson correlation coefficient comparing speed between Translator A and B for Dell is r=.19. For Autodesk it is slightly higher, r= .29. It appears the translators post-edited at quite different speeds even when initially presented with the same MT and source text.

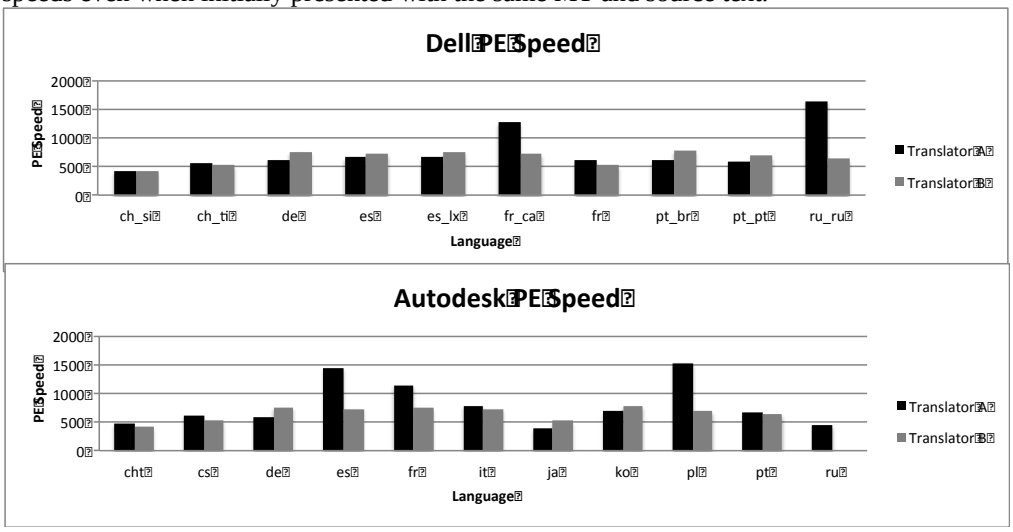


Figure 4 Autodesk and Dell PE speed per language for the same files.

The scatterplot in Figure 5 below shows the correlation between PE similarities as calculated using a character-based Levenshtein algorithm averaged across MT segments for each translator and PE speed in WPH in which self-review time was accounted for.

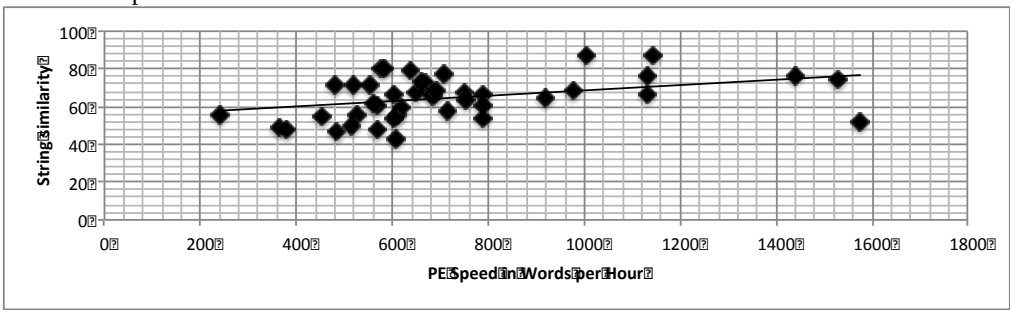


Figure 5. PE speed versus string similarity

44 translators across the Autodesk and Dell tests were bundled together in the analysis. A moderate correlation of Pearson r=3.7 was found. Thus far we have only discussed post-

editing speed. However, SLAB testing accounts for multiple segment categories so we turn our attention to HT.

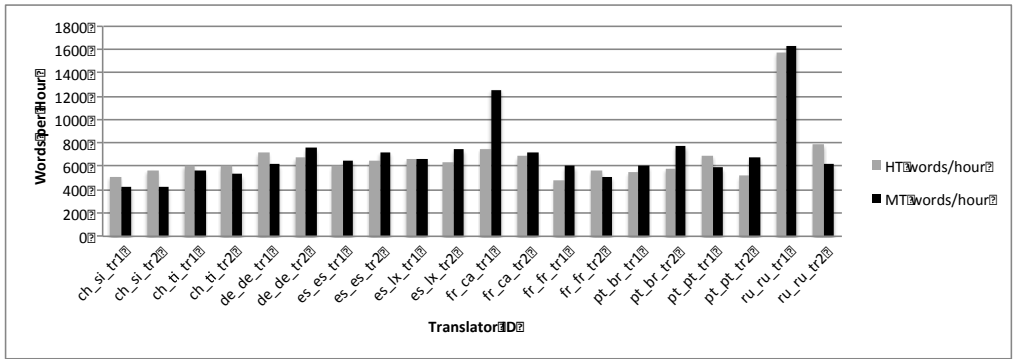


Figure 6. HT/MT speed ratios and absolute speed values for a Dell productivity test in 2012.

Figure 6 shows absolute HT and MT post-editing speeds for the Dell productivity test. As Autodesk had low HT sample sizes for languages except for French, Italian, German, Spanish (FIGS) it is not shown here. It should be noted that the +5% MT speed delta was positive mainly as a result of four efficient post-editors, fr_fr_tr1, pt_br_tr2 and pt_pt_tr2 and one very efficient post-editor fr_ca_tr1. Two years after this snapshot was taken, the current HT/MT speed ratio on the Dell account is around 40%. This improvement is due to improved MT but also translator selection weighted in favour of translators like these four translators and against their language counterparts. However, using string distance measures alone it is not always possible to identify efficient post-editors.

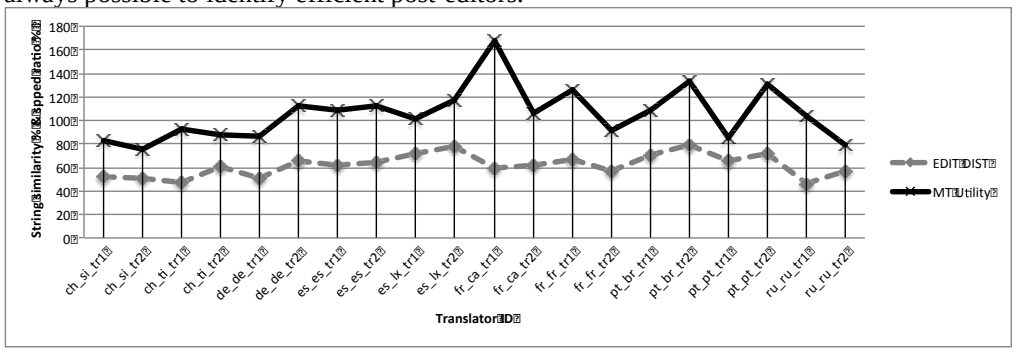


Figure 7. String similarity and MT utility (HT/MT speed ratio)

This can be seen in Figure 7 where the average string similarity across MT segments using character-based Levenshtein is shown per translator along with the HT/MT ratio. Both are represented as a line graph. For ease of visual comparison both string distance percentages (up to 80%) and PE speed ratio percentage (80% to 160%) are both shown on the X-axis. The three or four efficient posteditors are clearly seen as peaks above the 100% line, the breake-ven point for MT productivity. However, they are barely visible on the lower dotted line. In several instances edit distance is misleading as a predictor for PE speed relative to HT. For example, fr_ca_tr1 and fr_ca_tr2 have about the same edit distance value of 60% but very different PE speeds relative to HT. fr_ca_tr1 is twice as fast using MT relative to HT com-

pared to fr_ca_tr2. This is consistent with key findings from a similar large scale productivity test carried out by Plitt & Masselot (2010). They found the translator who post-edited fastest had the best quality review scores and also made the most changes. It seems some translators are more efficient that others at using MT to improve their working speed. More information on the MT system used by Autodesk can be found in Zhechev (2012) and a discussion of trends across multiple productivity tests at Welocalize is presented in by Casanellas & Marg (2014).

We conclude this discussion with a note on sample sizes. Though work remains to be done here, we have found that analyzing CAT-UAD over periods shorter than a day per main category (HT and MT) leads to unreliable statistics due to problems associated with small sample sizes and hence sparse data in each segment category. For this reason, it is important that productivity test costs are kept to a minimum. Once reliable speed data has been established, edit distance metrics can be used to monitor MT quality. However, a better solution would be to be able to measure post-editing speed at all times or at least intermittently in a translator's preferred desktop-based CAT tool.

5. Towards a standard for CAT-UAD in paid CAT tools

Though we are unaware of any research on the topic intuitively, translators are most familiar and hence very likely most productive in the one or two CAT tools they use most frequently. For the translators who responded to the questionnaire in Figure 1, in most cases this is Trados, Wordfast, MemoQ or DejaVu. OmegaT is only a standard CAT tool for a minority of translators.

However, HT/MT SLAB testing is useful to translators. In our experience the discounts for post-editing it facilitates are generally considered fair. A portion of the cost saved because of the time saved by MT can be passed on to the client. In turn this saving can be used to improve MT. As long as translator hourly earnings for MT post-editing are above hourly earnings for HT, payment for post-editing should not be a complaint. However, there are other reasons why temporal PE data is useful. Along with translator feedback, it also makes it possible to assign priority to problems reported by translators that impact most on working speed, e.g. over translation of URLs.

However, from a translator's perspective there is a darker side to measuring translation speed. It could be used to negotiate discounts where MT is not in play. For example, a translator who translates at 2,000 words per day when he is new to a regular translation account might translate at 3,000 words per day after a few months. Although they are paid by the word translators typically invest more time in terminology research when they are new to an account. Were a client to approach the translator with speed data to ask for a discount, at this point the translator is likely to become unhappy that translation speed data was being recorded as this time investment would be wasted.

In web-based CAT tools translation speed data can be recorded on a server and a translator cannot delete or view that data. In iOmegaT it is recorded on a local disk and can be deleted from the /instrumentation folder within the OmegaT project folder. In this respect desktop-based CAT tools provide a level of privacy web-based tools do not. However, a translator does not have to make a buying decision to use iOmegaT. Agencies are within their rights to ask freelance translators to use specific tools if they provide the tool at no cost. Equally, freelancer translators are within their rights to turn down the work.

The situation is different for paid desktop-based CAT tools like Trados, MemoQ or DejaVu. Freelance translators typically pay for this software. Were these applications to implement instrumentation of the kind used in iOmegaT, in our opinion it is important that speed

data is only shared at the discretion of the translator so it can be limited to post-editing or other scenarios in which technology can improve a translator's per hour earnings. Otherwise, it seems likely translators will pay for CAT tools that do not automatically share speed data. We do not see SLAB tests putting an end to per word pricing. However, we do feel that where an evolving technology like MT is concerned it is important for buyers who supply that MT and translators who use it to be able to monitor relative per hour earnings.

Finally, CAT-UAD may be useful beyond MT. It could be used by translators to communicate delays caused by slow servers (terminology and translation memory), time wasted fixing false positives generated by automatic translation quality checks or to measure the impact of technologies like intelligent fragment assembly, Automatic Speech Recognition (ASR) or predictive typing /auto-complete. It might also encourage training in the use of these technologies by measuring the impact of that training. As desktop-based and some web-based CAT tools become more complex it will become increasingly important to be able to record and analyze the impact of individual technologies on translation speed. We feel that if paid CAT tool companies address translator privacy concerns many translators will see the benefit of CAT-UAD and those that do not can ignore the feature.

To this end we have offered to provide early access to the iOmegaT CAT tool to a number of commercial desktop-based CAT tool publishers with a view to encouraging a discussion on CAT-UAD as a standard data format (Moran, 2014). Discussions are progressing in that regard.

6. Summary

In this paper we surveyed a number of applications that can be used to measure translation speed. We showed on the basis of a questionnaire that the desktop-architecture remains dominant in the CAT tool market and outlined some advantages of OmegaT, a free open-source CAT tool over web-based CAT tools. We briefly introduced the iOmegaT Translator Productivity Workbench and described how it is used in a typical enterprise translation workflow. On the basis of two HT/MT SLAB tests we showed that because translators vary greatly in how much MT speeds up their work it is important to be able to spot post-editors who are faster using MT and, conversely, recognize translators who are not aided by MT proposals. Finally, we made a case for instrumentation in existing paid CAT tools while drawing attention to the importance of a translator-centric privacy model.

Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of CNGL, the Centre for Global Intelligent Content (www.cngl.ie) at Trinity College Dublin and as part of Technology and an Innovation Development Award Feasibility Grant (12/TIDA/I2424) titled "iOmegaT – An Instrumented Replayable Computer-Aided-Translation Tool". The iOmegaT Workbench developed under this grant has been licensed to a small number of companies. Companies interested in licensing it or researchers interested in using it at no cost should contact John Moran at the e-mail address in the title of this paper.

References

Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., Garcia-Martinez, M., … others. (2013). Advanced Computer Aided Translation with a Web-Based Workbench. *Machine Translation Summit XIV*, 55.

Aziz, W., Castilho, S., & Specia, L. (2012). PET: a Tool for Post-editing and Assessing Machine Translation. In N. C. (Conference Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Bota, L., Schneider, C., & Way, A. (2013). COACH: Designing a new CAT Tool with Translator Interaction. In *Machine Translation Summit XIV, Main Conference Proceedings* (pp. 313–320).

Carl, M. (2012). Translog-II: a Program for Recording User Activity Data for Empirical Reading and Writing Research. In *LREC* (pp. 4108–4112).

Casanellas, L. &, & Marg, L. (2014). Assumptions, Expectations and Outliers in Post-Editing. In *EAMT 2014: Proceedings of the 17th Annual conference of the European Association for Machine Translation*.

Denkowski, M., & Lavie, A. (2012). TransCenter: Web-Based Translation Research Suite. In *Workshop on Post-Editing Technology and Practice Demo Session*. San Diego. Retrieved from http://www.cs.cmu.edu/~mdenkows/transcenter/

Depraetere, I., De Sutter, N., & Tezcan, A. (2014). Post-edited quality, post-editing behaviour and human evaluation: a case study. In S. O'Brien, M. Carl, L. Specia, & L. Balling (Eds.), *Post-editing of Machine Translation* (pp. 78–109). Newcastle upon Tyne.

Garcia, I. (2008). Power shifts in web-based translation memory. *Machine Translation*, *21*(1), 55–68. doi:10.1007/s10590-008-9033-6

Jakobsen, A. L., & Schou, L. (1999). Translog documentation. *Copenhagen Studies in Language*, *24*, 149–184.

Krings, H. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. The Kent State University Press.

Lewis, D., & Moran, J. (2010). Unobtrusive methods for low-cost manual evaluation of machine translation. In *Tralogy*. Paris.

Lossner, K. (2012). *memoQ 6 in Quick Steps*. Self published.

Moran, J. (2014). User Activity Data in a CAT tool - An industrial use case and a proposal for an open standard - oral presentation at FEISGILTT 2014. Dublin. Retrieved from http://www.localizationworld.com/lwdub2014/feisgiltt/accepted.html

Moran, J., Lewis, D., & Saam, C. (2014). Analysis of Post-editing Data: A Productivity Field Test using an Instrumented CAT Tool. In S. O'Brien, M. Carl, L. Specia, & L. Balling (Eds.), *Post-editing of Machine Translation* (pp. 126–146). Newcastle upon Tyne.

Plitt, M., & Masselot, F. (2010). A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, (93), 7–16.

Pym, A., Grin, F., Sfreddo, C., & Chan, A. L. J. (2013). *The status of the translation profession in the European Union.* London: Anthem.

Silva, R. (2014). Integrating Post-Editing MT in a Professional Translation Workflow. In S. O'Brien, M. Carl, L. Specia, & L. Balling (Eds.), *Post-editing of Machine Translation* (pp. 24–51). Newcastle upon Tyne.

Way, A., Holden, K., Ball, L., & Wheeldon, G. (2011). SmartMATE: Online self-serve access to state-of-the-art SMT. In *Proceedings of the Third Joint EM+/CNGL Workshop "Bringing MT to the User: Research Meets Translators", JEC* (pp. 43–52).

Zhechev, V. (2012). Machine Translation Infrastructure and Post-editing Performance at Autodesk. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012).* San Diego.

Translation Quality in Post-Edited versus Human-Translated Segments: A Case Study

Elaine O'Curran elaine.ocurran@welocalize.com
Welocalize Inc., 6 Dundee Park, Andover, Massachusetts 01810, USA

Abstract

We analyze the linguistic quality results for a post-editing productivity test that contains a 3:1 ratio of post-edited segments versus human-translated segments, in order to assess if there is a difference in the final translation quality of each segment type and also to investigate the type of errors that are found in each segment type. Overall, we find that the human-translated segments contain more errors per word than the post-edited segments and although the error categories logged are similar across the two segment types, the most notable difference is that the number of stylistic errors in the human translations is 3 times higher than in the post-edited translations.

1. Introduction

We have come to expect that using machine translation (MT) in combination with human post-editing (MTPE) saves on time and cost when compared with human translation (HT). However, there remains a lot of fear in the industry that integrating MT into the translation workflow will lead to lower quality. In order to find out for ourselves, we performed a detailed analysis of the linguistic quality assessment (LQA) reports for a post-editing productivity test we had recently carried out that involved the languages Brazilian Portuguese, French and Spanish and 6 translators (two per locale).

In some cases at the request of a client or an internal team, we conduct productivity tests in order to evaluate the usability of the raw MT output for post-editing by human translators and to estimate productivity gains over human translation from scratch. These tests are carried out in iOmegaT[1], an instrumented version of the open source translation tool OmegaT. Productivity tests are typically carried out for 8 hours and involve 2 translators per local. Of these 8 hours, 1 hour is used for revision. This is an important step for translators, as it mirrors their usual approach. As a testing environment, iOmegaT minimizes the disruption to a post-editor's process by providing an interface similar to those most frequently used in everyday production. The translation kit contains a mix of MT and HT segments as well as the usual formatting features (e.g. tags), and MT segments can be distinguished from internally propagated fuzzy matches. However, the kits are usually prepared to contain as few fuzzies as possible in order to maximize the MT and HT words in the tests. Glossaries and translation memories can be provided in the tool, translators can revisit segments and there is a spellcheck and tag validation feature for the review phase. Following delivery, a test kit is sent for LQA to assess the quality against the prescribed benchmarks for the content type, domain and language. By running this additional check, we can be certain that productivity gains are valid, not occurring at the expense of quality.

[1] CNGL Invention Disclosure, option period triggered 22/06/2012

2. Related Work

Fiederer and O'Brien (2009) set out to investigate if post-edited MT output was necessarily of lower quality than human translation and found that the post-edited machine translated output was assessed to be of higher clarity and accuracy, while the human translations were assessed to be of better style. Findings by Guerberof (2010) suggest that translators produce higher quality when using machine translated output than when processing fuzzy matches from translation memories. In her experiment, the number of errors found in TM segments was 91% higher than in MT segments. MT segments, on the other hand, contained 26% more errors than in the HT segments.

Plitt & Masselot (2010) used 12 professional localisation translators in their study and reported that translation jobs contained a higher number of errors than post-editing jobs. The MT engine had been trained on a large amount of company data, as this was a study carried out by Autodesk, and this scenario is highly representative for our case study.

Garcia (2010) was surprised at the quality results in his comparative study, which showed that the MT passages populated by Google Translator Toolkit and subsequently human-edited were more favourably assessed by the reviewers in 33 of 56 cases. This suggests that translating by post-editing MT output may be advantageous (Garcia, 2010). However, it is important to note that Garcia used trainee translators is in his study, whereas all the others studies referenced here employed professionals.

From her analysis of several months' LQA data for 4 language pairs, gathered from both human-translated and post-edited content, Peruzzi (2013) concludes that "the main differences in quality and types of errors are found between languages rather than translation scenarios, and […] these differences may not only be caused by quality of MT, but also by different cultural and linguistic aspects" (2013). Peruzzi's evaluation was based on two different workflows – one including MTPE and one including HT. The current use case differs in that it is based on a workflow that contains a mix of MTPE and HT segments within the same test environment.

3. Methods

3.1. MT profile

The machine translations are provided by a statistical MT system that has been specially customized for the client content using translation memories and glossaries.

3.2. Translator profile

All six translators are familiar with the account that is being tested and have at least 5 years translation experience and with the exception of the two, very senior, Spanish resources, all have between 1 to 4 years of post-editing experience.

3.3. Content profile

The content translated and post-edited in the test kits is real User Assistance content from the Software Antivirus/Security Compliance domain.

3.4. Reviewer profile

Dedicated third party account reviewers performed the LQA on the productivity kits to check compliance with standard quality expectations for the account. It was a blind review, i.e. the reviewers were not aware if the segment origin was MTPE or HT.

3.5. Linguistic review method

Similar to the LISA QA Model and SAE J2450, our applied QA metrics are a quantitative-based method of translation quality assessment which measures the number, severity and type of errors found in a text and calculates a score based on the number of words reviewed, indicative of the quality of a given translation. The reviewers evaluate the translations based on the following criteria:

1. Accuracy: Cross References, Omission/Addition, Incorrect Translation/Meaning, Unlocalized Text
2. Language: Punctuation, Spelling/Typo, Grammar/Syntax
3. Terminology: Context, Inconsistency, Glossary
4. Style: General Style, Client Style Guide, Language Variants/Slang, Register/Tone, Unnecessary Additions
5. Country: Country/Regional Standards, Local Market Suitability
6. Functional: Format, Hidden Text, Tags/Links, Technical Procedures, Spacing

The severity levels Minor or Major are applied to each error, based on the definitions in Table 1 below.

Major errors are blatant and severe errors that jeopardize, inverse or distort the meaning of a translation. Major errors are severe failures in accuracy, compliance, or language. Examples:
– Any statement that can be potentially offensive.
– Errors that endanger the integrity of data or the health/safety of users.
– Errors that modify or misrepresent the functionality of the device or product.
– Errors that clearly show that the client's and/or Welocalize' instructions haven't been followed.
– Errors that appear in a High Visibility Portion and/or is numerously repeated.
– Grammar or syntax errors that are gross violations of generally accepted language conventions.
Minor errors are all errors that do not fall under major severity as defined above nor are merely preferential changes. Examples:
– Accuracy errors that result in a slight change in meaning.
– Small errors that would not confuse or mislead a user but could be noticed.
– Formatting errors not resulting in a loss of meaning, e.g. wrong use of bold or italics.
– Wrong use of punctuation or capitalization not resulting in a loss of meaning.
– Generic error to indicate generally inadequate style (e.g. literal translation, "stilted" style, etc.).
– Grammar or syntax errors that are minor violations of generally accepted language conventions.
– Typos and misspellings that do not result in a loss of meaning.

Table 1: Error Severity Descriptions

3.6. LQA results

For this productivity test, 3 of the kits returned a fail. This was one of the drivers behind this case study, as we wanted to understand if the underediting in these kits could be traced back to MT segments that had not been post-edited properly, or if the translators had simply not performed an adequate self-review in general. It is worth noting that the two inexperienced post-editors delivered the best overall quality.

Translator	LQA Result
ES - 1	PASS
ES - 2	PASS
FR - 1	PASS
FR - 2	FAIL
PTBR - 1	FAIL
PTBR - 2	FAIL

Table 2: LQA results per resource

The Pass treshold is 99.60% based on the following mathematical algorithm : =(1-(Minor_Errors+(2*Major_Errors*Major_Errors))/Sample_Size).

3.7. Review scope

Across three locales and 6 translators, approximately 8000 words were reviewed. Figure 1 below illustrates the exact number of words that were reviewed per locale and per segment type. The reason we see less words for Brazilian Portuguese is that the reviewers were instructed to spend 1 hour on each translator's kit and mark all the segments that had been checked. Due to the higher number of errors for this locale, the reviewer covered less words.

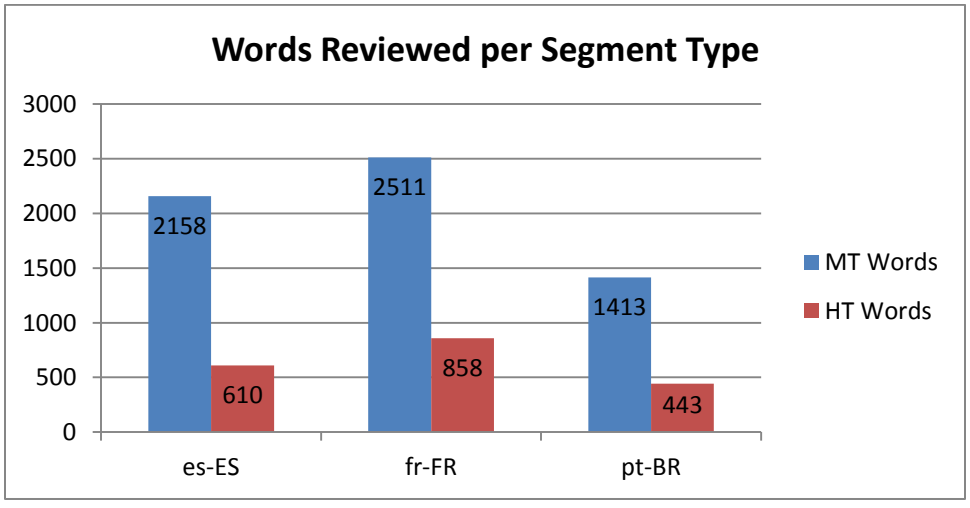


Figure 1: Words reviewed by locale and segment type

Figure 1 also illustrates the approximate 3:1 ratio of MT versus HT words in the reviewed kits.

4. Results

4.1. Errors per 1000 words

To account for the difference in word count per segment type, we calculate errors per 1000 words. As illustrated in Figure 2, we found that there were more errors in HT segments than in MT segments across all three locales. This consistent result was surprising considering the different levels of quality that had been delivered for this test.

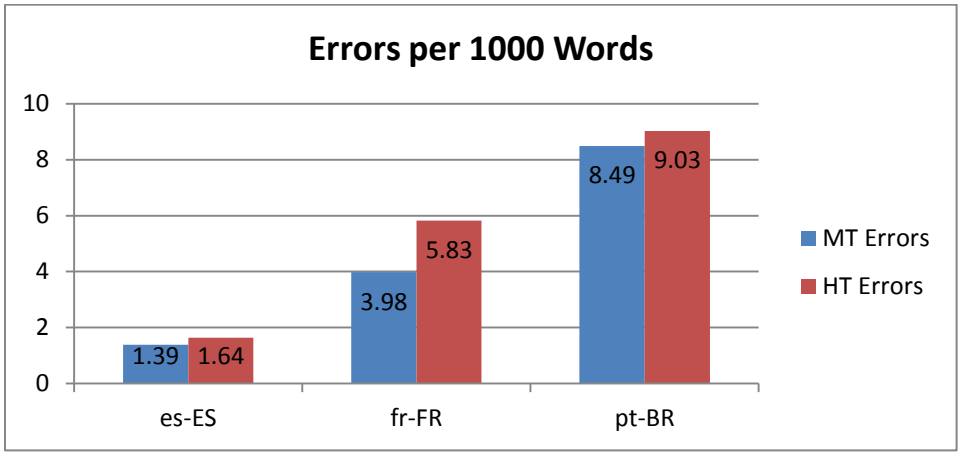


Figure 2: Errors per 1000 words

4.2. Error types found

For illustration purposes, the Table 3 error table below has been normalized to account for the 3:1 ratio of MTPE words versus HT words. The HT errors have simply been multiplied by three. We see that overall there are 25 errors in MTPE, including 5 Major errors, while there are 30 errors in HT including 6 Major errors. There are more Language, Style and Tag errors in HT, while MTPE has more Glossary errors. The Accuracy errors are quite evenly distributed: both have exactly 6 Minor and 3 Major errors in the Accuracy category. The most notable difference is that the number of stylistic errors in the HT segments is 3 times higher than in the MTPE segments.

Error Category	MTPE Errors	HT Errors
Accuracy - Meaning/Incorrect Translation - Major	1	0
Accuracy - Meaning/Incorrect Translation - Minor	3	6
Accuracy - Omissions/Additions - Major	2	3
Accuracy - Omissions/Additions - Minor	3	0
Functional - Tags/Links - Major	1	3
Functional - Tags/Links - Minor	1	0
Language - Grammar/Syntax - Major	1	0
Language - Grammar/Syntax - Minor	6	6

Language - Punctuation - Minor	0	3
Style – Client Style guide - Minor	1	0
Style - Language variants/slang - Minor	1	0
Style - General style - Minor	0	3
Style - Unnecessary Additions - Minor	0	3
Terminology - Glossary adherence - Minor	5	3
Total errors	**25 (5 major)**	**30 (6 major)**

Table 3: Error types in MTPE and HT

5. Conclusions

The result of this case study supports the findings of some of the other studies mentioned above that found fewer errors in MT post-edited work than in human translations. While there is some overlap between the types of errors found in the human-translated segments and post-edited segments, it is notable that more errors were found in human translations in categories such as Punctuation, Tags and Style. However, our case study was performed on relatively small volumes, the three locales are Romance languages and the content type is technical. In order to draw firm conclusions, it would be important to conduct a larger study with more diverse languages and content types and to also include fuzzy matching for additional benchmarking.

Acknowledgement

The author would like to acknowledge the support of Olga Beregovaya, Laura Casanellas and Lena Marg.

References

Fiederer, R. and O'Brien, S. (2009) Quality and Machine Translation: A realistic objective? *The Journal of Specialised Translation*, Issue 11. http://www.jostrans.org/issue11/art_fiederer_obrien.pdf. Accessed 17 July 2014

Garcia, I. (2010) Is machine translation ready yet? *Target*, 22(1): 7–21

Guerberof, A. (2009) Productivity and Quality in MT Post-Editing, Universitat Rovira I Virgilli, Spain. http://www.mt-archive.info/MTS-2009-Guerberof.pdf Accessed 17 July 2014.

Peruzzi, S. (2013) Investigating the Impact of Machine Translation and Post-Editing on Quality and Errors in a Translation Memory-based Workflow. Dissertation for the MSc in Translation Technology at DCU, Ireland

Plitt, M., and Masselot, F. (2010) A Productivity Test of Statistical machine Translation Post Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*. Jan. Num 93. 7-16.

TAUS Post-Editing course

(DEMO SUBMISSION)

Attila Görög (TAUS DQF product manager) – attila@taus.net

While there is a massive adoption of MT post-editing as a new service in the global translation industry, a common reference to skills and best practices to do this work well has been missing. TAUS took up the challenge to provide a course that would integrate with the DQF tools and the post-editing best practices developed by TAUS members in the previous years and offers both theory and practice to develop post-editing skills. The contribution of language service providers who are involved in MT and post-editing on a daily basis allowed TAUS to deliver fast on this industry need. This online course addresses the challenges for linguists and translators deciding to work on post-editing assignments and is aimed at those who want to learn the best practices and skills to become more efficient and proficient in the activity of post-editing.

The course contains six modules and can be completed in a week (approx. one hour per module, the evaluation and the post-editing assignments excluded). Upon successful completion, students obtain the TAUS Post-Editing Certificate and a stamp to add to their email signature and/or web site. The course consists of generic modules covering the general skills and practices. This material provides comprehensive background information on machine translation, useful details on facts, trends and good general knowledge about post-editing. It helps you understand the major challenges and advantages of machine translation. And it prepares participants to successfully manage machine-translation projects and to become confident post-editors. The course also contains two language-specific modules addressing the particular errors by different target languages. These practical modules contain an MT evaluation exercise and the post-editing of a real machine-translation output.

The ideal participants for this course are prepared individuals who have:

- Some experience in translation and editing
- Interest in learning more about machine translation and post-editing
- A basic knowledge of translation technology
- Perform at native or near-native level and have good translation skills in the selected language.

This course is designed for a wider audience within the language industry, including, but not exclusively for:

- Translation students participating in translation training programmes at a university or college
- Active translators and editors interested in broadening their skill-set
- Teachers and researchers who are involved in translator training and are interested in this new area of the industry
- Employees working in different positions at linguistic services providers (project managers, terminologists and language technologists)
- Anyone who is interested in translation automation

The post-editing course is a collaborative industry initiative coordinated by TAUS in close cooperation with companies such as Welocalize, Hunnect, CPSL, Arabize, Version Internationale, Global Textware and other LSPs. At the launch (in April 2014), the course was available with *Spanish, French, Arabic, Hungarian* and *Dutch* language-specific modules. In the summer of 2014, new modules are being added including *Japanese, Korean, Turkish, Polish, German* and *Italian*. More language service providers will start collaborating on this project as more languages will be added in the remainder of this year and next year.

To promote the course, TAUS has been organizing free webinars on Post-editing. This webinar series introduces post-editing as an emerging profession within the translation industry. Participants discuss best practices and offer an overview of available tools and methods for post-editing. Each webinar has a language focus and invited speakers elaborate on post-editing problems of the given language. For more information, please refer to: https://www.taus.net/taus-post-editing-webinar.

The TAUS post-editing course is available on a subscription basis. Special prices are available for academic institutions. For more information, please refer to: https://evaluation.taus.net/post-editing-course-pricing.

This demo will offer participants the chance to go through selected modules of the Post-editing course, to get an impression of the language-specific assignments and to ask questions.

TAUS Post-editing productivity tool

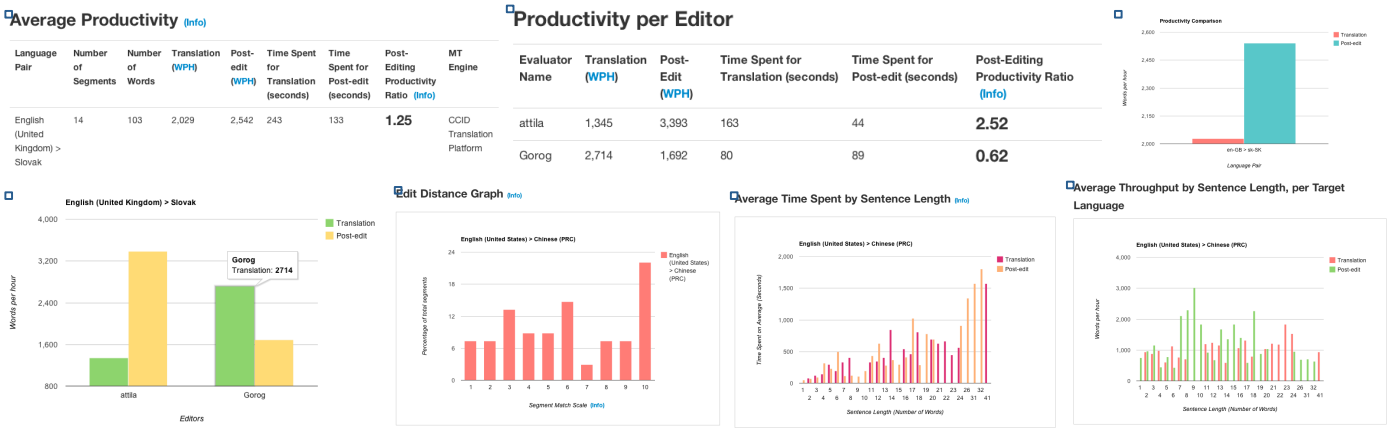
(DEMO SUBMISSION)

Attila Görög (TAUS DQF product manager) – attila@taus.net

The online TAUS QE platform (DQF) provides three types of quality evaluation methods: 1. Quality Assessment based on adequacy, fluency and error-typology; 2. Comparison and ranking of MT engines and 3. Post-editing productivity testing. The platform offers a tool neutral and vendor independent environment for the human evaluation of translation quality. Users gather vital data to help establish return-on-investment, measure productivity enhancements, and benchmark performance, helping to ensure informed decisions are taken.

In this demo session, we will focus on the third type of evaluation method: post-editing productivity testing. The tool enables project managers to either measure post-editing productivity by asking post-editors to post-edit the full MT-output or to compare translation to post-editing by letting them translate half of the text from scratch and post-edit the other half. The first type of testing provides information on Time to edit (average number of words processed in a given timespan) and Post-editing effort (average percentage of word changes applied) while the second type enables comparison of the difference in speed between MT post-editing and translation from scratch by measuring time and edit distance.

After the productivity testing is completed, the reporting tool provides information on both the average productivity and the productivity per post-editor including the number of words and segments processed, the number of words post-edited per hour, the translated words per hour (if relevant), the time spent for post-editing, the time spent for translation in seconds (if relevant), the productivity ratio score (when post-editing is compared to translation), the average time spent by sentence length, the average throughput by sentence length, per target language and the edit distance graph.



Average Productivity (Info)

Language Pair	Number of Segments	Number of Words	Translation (WPH)	Post-edit (WPH)	Time Spent for Translation (seconds)	Time Spent for Post-edit (seconds)	Post-Editing Productivity Ratio (Info)	MT Engine
English (United Kingdom) > Slovak	14	103	2,029	2,542	243	133	1.25	CCID Translation Platform

Productivity per Editor

Evaluator Name	Translation (WPH)	Post-Edit (WPH)	Time Spent for Translation (seconds)	Time Spent for Post-edit (seconds)	Post-Editing Productivity Ratio (Info)
attila	1,345	3,393	163	44	2.52
Gorog	2,714	1,692	80	89	0.62

One can also use the tool to compare the output of different MT engines, to try to understand the impact of certain content issues (e.g. terminology, length etc.), the variance across content types, the correlation with certain metrics and scores or the influence of the translator profile (age, gender, experience), etc. In combination with other tools provided by the QE platform, one can correlate indirect methods (post-editing productivity) with direct methods (error-typology or adequacy/fluency).

As of February 2014, the TAUS QE platform is free for academic use and is available on a subscription basis for commercial use. For more information, please refer to: https://evaluation.taus.net/

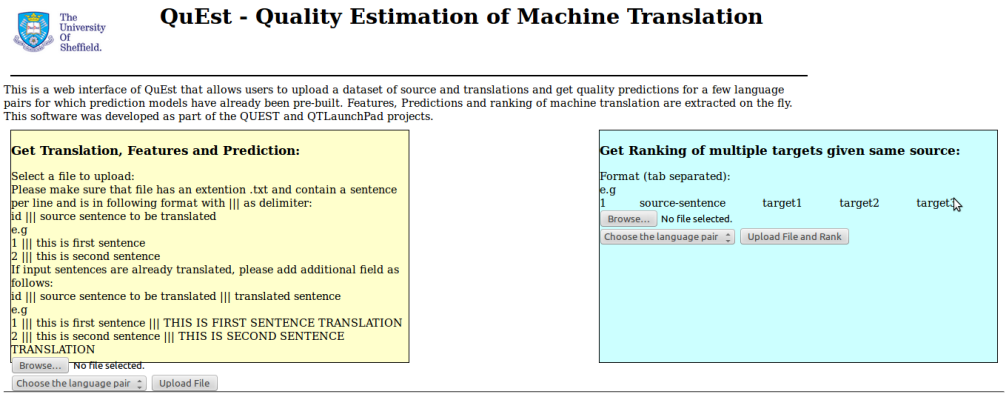
This demo will offer participants the chance to try out the TAUS QE platform including the post-editing productivity tool, to view sample reports and to ask questions.

QUEST: a framework for translation quality estimation

Lucia Specia l.specia@sheffield.ac.uk
Kashif Shah kashif.shah@sheffield.ac.uk
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street, S4 1DP, UK

We present QUEST, an open source framework for translation quality estimation. QUEST provides a wide range of feature extractors from source and translation texts and external resources and tools. These go from simple, language-independent features, to advanced, linguistically motivated features. They include features that rely on information from the translation system and features that are oblivious to the way translations were produced. In addition, it provides wrappers for a well-known machine learning toolkit, `scikit-learn`,[1] including techniques for feature selection and model building, as well as parameter optimisation.

We also present a Web interface and functionalities for non-expert users. Using this interface, quality predictions (or internal features of the framework) can be obtained without the installation of the toolkit and the building of prediction models. The interface also provides a ranking method for multiple translations given for the same source text according to their predicted quality.



Acknowledgements

This work was supported by funding from the from European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 296347 (QTLaunchPad).

[1] `http://scikit-learn.org/`

An Open Source Desktop Post-Editing Tool

Lane Schwartz lanes@illinois.edu
Department of Linguistics, University of Illinois at Urbana-Champaign,
Foreign Languages Building, 707 S Mathews Ave, Urbana, IL 61801, USA

We present a simple user interface for post-editing that presents the user with the source sentence, machine translation, and word alignments for each sentence in a test document (Figure 1). This software is open source, written in Java, and has no external dependencies; it can be run on Linux, Mac OS X, and Windows.

This software was originally designed for monolingual post-editors, but should be equally usable by bilingual post-editors. While it may seem counter-intuitive to present monolingual post-editors with the source sentence, we found that the presence of alignment links between source words and target words can in fact aid a monolingual post-editor, especially with regard to correcting word order. For example, in our experiments using this interface (Schwartz et al., 2014), post-editors encountered some sentences where a word or phrase was enclosed within bracketing punctuation marks (such as quotation marks, commas, or parentheses) in the source sentence, and the machine translation system incorrectly reordered the word or phrase outside the enclosing punctuation; by examining the alignment links the post-editors were able to correct such reordering mistakes.

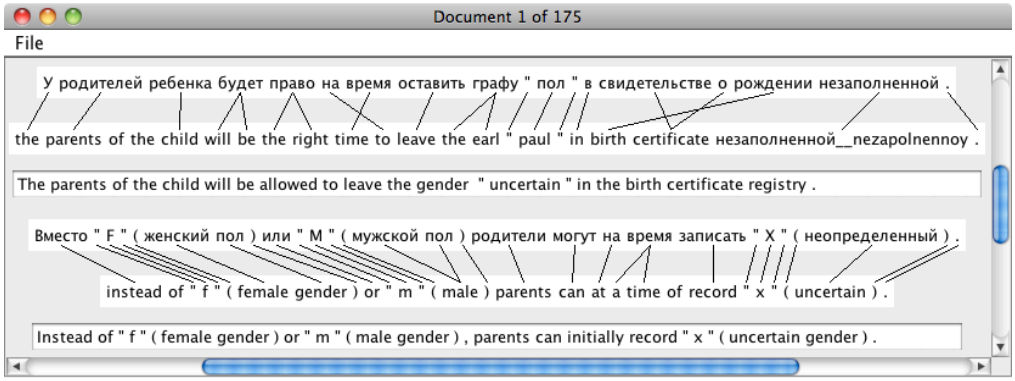


Figure 1: Post-editor user interface

References

Schwartz, L., Anderson, T., Gwinnup, J., and Young, K. M. (2014). Machine translation and monolingual postediting: The AFRL WMT-14 system. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 186–194, Baltimore, Maryland. Association for Computational Linguistics.

Real Time Adaptive Machine Translation: cdec and TransCenter

Michael Denkowski Alon Lavie Isabel Lacruz* Chris Dyer
Carnegie Mellon University, Pittsburgh, PA *Kent State University, Kent, OH
{mdenkows,alavie,cdyer}@cs.cmu.edu ilacruz@kent.edu

cdec Realtime and TransCenter provide an end-to-end experimental setup for machine translation post-editing research. Realtime[1] provides a framework for building adaptive MT systems that learn from post-editor feedback while TransCenter[2] incorporates a web-based translation interface that connects users to these systems and logs post-editing activity. This combination allows the straightforward deployment of MT systems specifically for post-editing and analysis of translator productivity when working with adaptive systems. Both toolkits are freely available under open source licenses.

1 Adaptive Machine Translation with cdec

In contrast to traditional machine translation systems that operate in batch mode, cdec Realtime immediately incorporates post-editor feedback during translation tasks. Three major MT system components are extended to support online updates, allowing new data to be incorporated after each sentence is translated:

- An online translation model is updated to include new translations extracted from post-editing data.
- A dynamic language model is updated to include post-edited target language text.
- An online update is made to the system's feature weights after each sentence is post-edited.

Live post-editing experiments have shown that these extensions result in translations that require less effort to post-edit and are preferred by human translators.

A Realtime system operates as follows. Single instances of the large initial translation and language models are loaded into memory. When a new user requests a translation, a new *context* is started that includes user-specific dynamic translation and language models plus a decoder instance with user-specific weights. When a sentence is translated, the user-appropriate decoder combines the initial and user-specific models. When a post-edited sentence is available as feedback, the following happen in order: (1) the source-reference pair is used to update feature weights, (2) translation rules from the source-reference pair are added to the user-specific translation model, and (3) the user-specific language model is updated with the reference. In the latest version of Realtime, weight updating during both optimization and decoding replaces the standard BLEU metric with a version of Meteor specifically targeting post-editing, yielding further reduction in translator effort.

2 Data Collection and Analysis with TransCenter

TransCenter provides a web-based translation editing interface that remotely records user activity. Translators use a web browser to access a familiar two-column editing environment. The left column displays the source sentences while the right column is incrementally populated with translations from a Realtime system as the user works. During editing, all key presses and mouse clicks are logged so that the full editing process can be analyzed. As each sentence is edited, the resulting translation is reported to the Realtime system for learning and the next translation is generated. The user is additionally asked to rate the amount of work required to post-edit each sentence immediately after completing it, yielding maximally accurate feedback. TransCenter also records the number of seconds each sentence is focused and provides a pause button for when translators needs to take breaks. TransCenter can generate reports of translator effort as measured by (1) keystroke, (2) exact timing, and (3) actual translator post-assessment. Final translations and millisecond-level timings of every user action are available for further analysis.

[1] http://www.cs.cmu.edu/~mdenkows/cdec-realtime.html
[2] http://github.com/mjdenkowski/transcenter-live

Post-editing User Interface Using Visualization of a Sentence Structure

Yudai Kishimoto† **Toshiaki Nakazawa**‡ **Daisuke Kawahara**† **Sadao Kurohashi**†
†Graduate School of Informatics, Kyoto University, Kyoto 606-8501
‡Japan Science and Technology Agency, Kawaguchi-shi, Saitama 332-0012
† {kishimoto,dk,kuro}@nlp.ist.i.kyoto-u.ac.jp
‡ nakazawa@pa.jst.jp

1 Purpose and characteristics

Translation has become increasingly important by virtue of globalization. To reduce the cost of translation, it is necessary to use machine translation and further to take advantage of post-editing based on the result of a machine translation for accurate information dissemination. Such post-editing (e.g., PET [Aziz et al., 2012]) can be used practically for translation between European languages, which has a high performance in statistical machine translation. However, due to the low accuracy of machine translation between languages with different word order, such as Japanese-English and Japanese-Chinese, post-editing has not been used actively.

We propose a post-editing system based on syntax-based machine translation to deal with different word order. For language pairs with different word order, it is time-consuming for a translator to understand what a machine translation system did. To solve this problem, our system displays the following three portions: a parse of the source language (A), a translation that keeps the word order of the source language (B), and a translation in the word order of the target language (C). This visualization makes the translator efficiently evaluate the quality of the translations and flexibly use the translations in various levels as follows.

1. If the parse or the translation is disorganized, the translator gives up using it and translates the source sentence from scratch. This can be efficiently judged mainly from A. Previous post-editing systems only displayed a final translation and made this judgment difficult.

2. If the translation is partially correct but has errors in word order, the translator changes the word order based on B. To make this process efficient, B is editable and translation blocks can be swapped on GUI.

3. If the translation does not have major errors including errors in word order, the translator makes a few revisions based on C. C is also editable and translation blocks can be swapped.

2 System description

The input of our system is a text file which contains the parse result of an original sentence and a translation result and a translation mapping.[1] This system only uses part of speech tags and dependency relations and can be expanded to many languages if we prepare these data. The output is a JavaScript file, and we can view the system result on a Web browser as in Figure 1.

Figure 1: Sample of post-editing interface

In Figure 1, we display three rectangles explained in section 1: A as a rectangle whose background color is sky blue B as a rectangle whose background color is pink, and C as a rectangle whose background color is orange. We can edit these rectangles in the way described in section 1.

3 Conclusion

We present a post-editing user interface using visualization of sentence structure. This system helps us to analyze the cause of errors more easily and hopefully will improve the efficiency of post-editing.

References

Aziz, W., Castilho, S., and Specia, L. (2012). Pet: a tool for post-editing and assessing machine translation. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3982–3987, Istanbul, Turkey. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1587.

[1]Sample text is available at `http://lotus.kuee.kyoto-u.ac.jp/~yudaik/zh-ja_sample.txt`

Kanjingo: A Mobile App for Post-Editing

Sharon O'Brien **Joss Moorkens** **Joris Vreeke**
CNGL: Centre for Global Intelligent Content, Dublin City University

We present Kanjingo, a mobile app for post-editing currently running under iOS. The App was developed using an agile methodology at CNGL, DCU. Though it could be used for numerous scenarios, our test scenario involved the post-editing of machine translated sample content for the non-profit translation organization Translators Without Borders. Feedback from a first round of user testing for English-French and English-Spanish was positive, but users also identified a number of usability issues that required improvement. These issues were addressed in a second development round and a second usability evaluation was carried out in collaboration with another non-profit translation organization, The Rosetta Foundation, again with French and Spanish as target languages.

As post-editing is a demanding text manipulation task, it was assumed that the limitations of the mobile environment may make this task unworkable. However, feedback from our two user groups has been positive and demonstrates that the Kanjingo layout and functionality makes mobile post-editing feasible.

We do not propose that this should replace desktop tools for post-editing. However, especially in volunteer scenarios where resources and time are limited, we envisage volunteers post-editing small segments of text as they find snippets of time in their daily schedules (e.g. waiting for the bus or train to arrive).

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

Figure 1: Kanjingo Segment Selection Screen

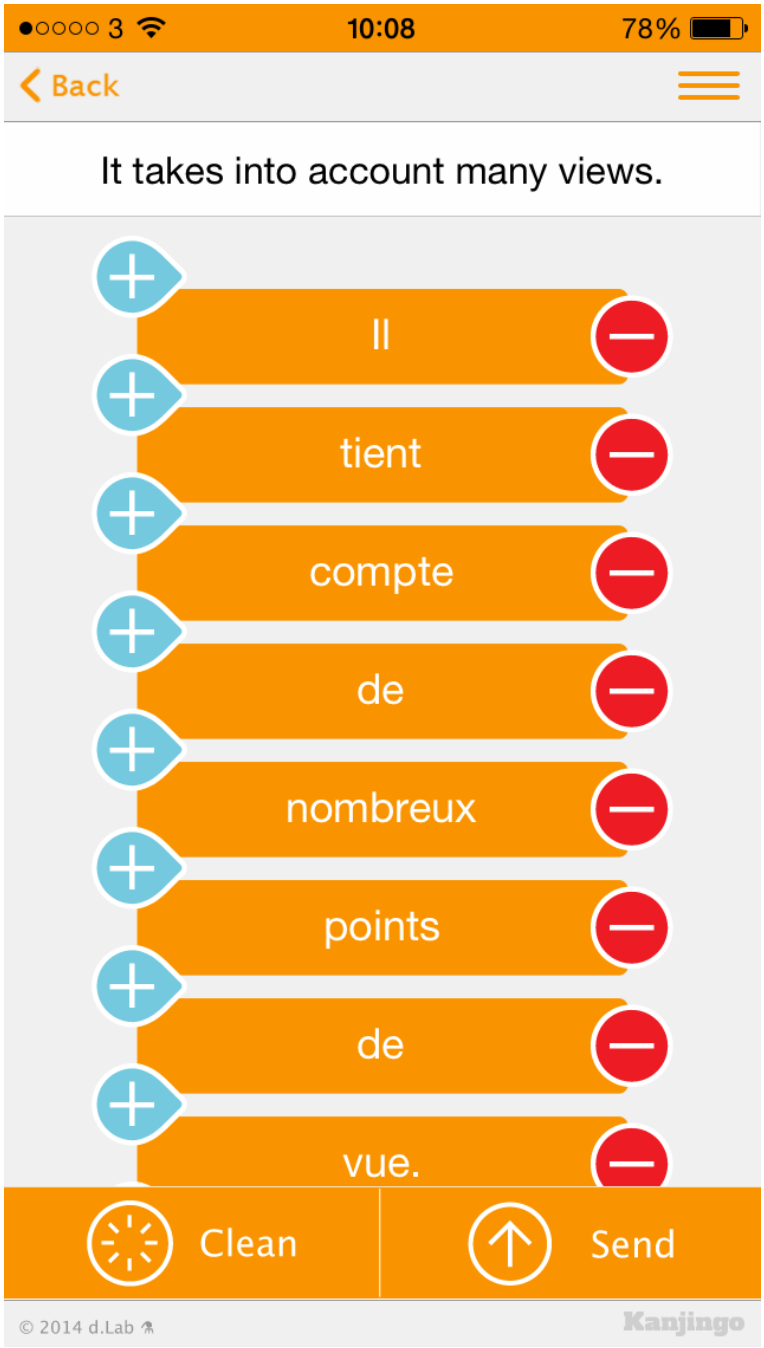


Figure 2: Kanjingo edit screen