Machine Translation Summit XV
http://www.amtaweb.org/mt-summit-xv

WPTP 2015:
Post-editing Technology and Practice

Organizers:
Sharon O’Brien (Dublin City University)
Michel Simard (National Research Council Canada)
Joss Moorkens (Dublin City University)
Introduction

The fourth Workshop on Post-Editing Technology and Practice (WPTP4) was organised for November 3rd, 2015, in Miami, USA, as a workshop of the XVth MT Summit. This was the fourth in a series of workshops organised since 2012 (WPTP1 – San Diego 2012, WPTP2 – Nice 2013, WPTP3 – Vancouver 2014).

The accepted papers for WPTP4 cover a range of topics such as the teaching of post-editing skills, measuring the cognitive effort of post-editing, and quality assessment of post-edited output. The papers were also book-ended by two invited talks by Elliott Macklovitch (Independent Consultant) and John Tinsley (Iconic Translation Machines). Elliott Macklovitch’s talk was entitled “What Translators Need to Become Happy Post-Editors” while John Tinsley’s talk was entitled “What MT Developers Are Doing to Make Post-Editors Happy”. By juxtaposing these two points of view we hoped to provide an interesting frame for attendees where the views of users and developers were represented.

We sincerely thank the authors of the papers as well as our two invited speakers for sharing their research. Thanks are also extended to the Programme Committee for their detailed reviews. We also thank Dr. Joss Moorkens (DCU) for his work with the proceedings, as well as the MT Summit organising committee for their support throughout.

The Program Committee of the Workshop on Post-Editing Technology and Practice (WPTP4)

Michel Simard
Sharon O'Brien
# Contents

4th Workshop on Post-Editing Technology and Practice (WPTP4)

1 Introduction  
Sharon O’Brien, Michel Simard  

2 How to Teach Machine Translation Post-Editing? Experiences from a Post-Editing Course  
Maarit Koponen  

16 Quality Assessment of Post-Edited versus Translated Wildlife Documentary Films: A Three-Level Approach  
Carla Ortiz Boix, Anna Matamala  

31 The Impact of Machine Translation Error Types on Post-Editing Effort Indicators  
Joke Daems, Sonia Vandepitte, Robert Hartsuiker, Lieve Macken  

46 Living on the Edge: Productivity Gain Thresholds in Machine Translation Evaluation Metrics  
Carla Parra Escartín, Manuel Arcedillo  

57 Domain Adaptation for Social Localisation-based SMT: A Case Study Using Trommons Platform  
Jinhua Du, Andy Way, Zhengwei Qiu, Asanka Wasala, Reinhard Schaler
Introduction

The fourth Workshop on Post-Editing Technology and Practice (WPTP4) was organised for November 3rd, 2015, in Miami, USA, as a workshop of the XVth MT Summit. This was the fourth in a series of workshops organised since 2012 (WPTP1 – San Diego 2012, WPTP2 – Nice 2013, WPTP3 – Vancouver 2014).

The accepted papers for WPTP4 cover a range of topics such as the teaching of post-editing skills, measuring the cognitive effort of post-editing, and quality assessment of post-edited output. The papers were also book-ended by two invited talks by Elliott Macklovitch (Independent Consultant) and John Tinsley (Iconic Translation Machines). Elliott Macklovitch’s talk was entitled “What Translators Need to Become Happy Post-Editors” while John Tinsley’s talk was entitled “What MT Developers Are Doing to Make Post-Editors Happy”. By juxtaposing these two points of view we hoped to provide an interesting frame for attendees where the views of users and developers were represented.

We sincerely thank the authors of the papers as well as our two invited speakers for sharing their research. Thanks are also extended to the Programme Committee for their detailed reviews. We also thank Dr. Joss Moorkens (DCU) for his work with the proceedings, as well as the MT Summit organising committee for their support throughout.

The Program Committee of the Workshop on Post-Editing Technology and Practice (WPTP4)

Michel Simard
Sharon O’Brien
How to teach machine translation post-editing?
Experiences from a post-editing course

Maarit Koponen  maarit.koponen@helsinki.fi
Department of Modern Languages, University of Helsinki, Helsinki, Finland

Abstract
Advances of machine translation technology have in recent years increased its use in various contexts. In the translation industry, work processes involving the use of machine translated texts as a raw translation to be post-edited by a translator are becoming increasingly common. The increasing use of post-editing processes also raises questions related to teaching and training of translators and post-editors, and institutions offering translator training have started to incorporate post-editing in their curricula. This paper describes a machine translation and post-editing course arranged at the University of Helsinki in Fall 2014. From the teacher’s perspective, we discuss experiences of planning and teaching of the post-editing course. The development of the students’ experiences and perception of the course contents, machine translation technology and post-editing are also discussed based on reflective essays written by the students after the course.

1 Introduction
The development of machine translation (MT) quality has led to growing use of the technology in recent years in many contexts. In the professional context, an increasingly common workflow involves the use of machine translated text as a raw translation to be corrected or post-edited by a translator. Tools and practices for such workflows are being developed in large multilingual organizations, such as the European Commission (Bonet, 2013), and by language service providers who increasingly offer their clients tailored MT systems and post-editing (PE) services. Based on a recent survey of 438 stakeholders in the translation and localisation field, Gaspari et al. (2015) report that 30% currently use MT, and the majority (70%) of the MT users combine that with PE at least some of the time. The use and usability of MT and PE varies greatly in different countries and language pairs, however.

The increasing use of MT and PE workflows in the translation industry has also attracted interest to how translators and post-editors are, or should be, trained. The need for specific training for post-editing has been argued by O’Brien (2002), for example. More recently, based on a review of various translation industry surveys in addition to their own survey, Gaspari et al. (2015) argue that there is a growing demand for MT services as well as a growing demand for expertise in PE skills, and emphasize the impact of “familiarity with translation technology” on the employability of future translators.

To answer to this changing industry landscape, universities training translators have therefore started to incorporate MT and PE in their curricula. In this paper, we describe experiences from a special course focusing on post-editing, arranged at the University of Helsinki, Finland, during Fall term 2014. The aim of the course was to introduce translator students to the theory and practice of MT and PE, covering topics ranging from the technical principles of MT to
PE practice, and research related to PE processes and MT evaluation. This paper presents an overview of the course contents, learning outcomes and the lecturer’s experiences from planning and teaching the course. The students’ perspectives are also discussed based on reflective essays written after the course.

The remainder of the paper is arranged as follows: Section 2 presents a brief overview of the current status of MT and PE use in Finland to provide context for the course described. Section 3 presents related work on MT and PE in translator training. Section 4 provides an overview of the course itself and an evaluation of the practical organization and learning outcomes. Section 5 discusses the themes arising from the students’ reflective essays as well as issues raised during the course. Finally, Section 6 presents the conclusions and some recommendations for future teaching.

2 MT and PE use in Finland

The adoption rate of MT and PE processes naturally varies in different countries and language pairs. For some language pairs, such as English-Spanish, MT systems have already for some time been able to provide sufficient quality to make PE a viable alternative. For some other language pairs, however, MT quality has lagged behind and PE has not been seen as feasible.

Finland is one of the countries where the use of MT and PE has not been particularly widespread. In large part, the situation is connected to MT quality achievable with Finnish as one part of the language pair. As a morphologically rich language with relatively free word order, Finnish has proven difficult for MT systems. For a more detailed discussion of the particularities of the Finnish language related to MT, see Koskenniemi et al. (2012). The quality issues are reflected, for example, in the European Commission trials, where English-Finnish translators considered MT at most sufficient to suggest ideas for expressions or not usable at all and better replaced by translation from scratch (Leal Fontes, 2013). In contrast, for more successful language pairs like French-Spanish, French-Italian and French-Portuguese, most MT segments were rated reusable.

On the other hand, also the small market area has attracted relatively little interest in developing MT systems for Finnish. In addition to free online statistical machine translation systems like Google Translate, the survey by Koskenniemi et al. (2012) mentions only two rule-based machine translation R&D projects, one of which did not reach product stage. A recent survey of 238 Finnish translators found that MT-related skills were not considered important by the translators themselves, which is likely connected to the unavailability and low quality of MT systems involving Finnish (Mikhailov, 2015). The report of this survey does not provide exact numbers on how many respondents used MT systems, but noted that most were only familiar with free online systems, with only three having used systems other than Google (Mikhailov, 2015, p. 111, endnote 5). Development work is, however, reportedly being carried out by Finnish language service providers.

3 MT and PE in translator training

As the use of MT and PE workflows has increased, research on, and teaching of, skills specific to post-editing has become necessary. Part of the skill set is likely to be shared with “traditional” human translation, such as source and target language proficiency, subject area knowledge, text linguistic skills, cultural and intercultural competence, as well as general documentation and research skills (see O’Brien, 2002; Rico and Torrejón, 2012; Austermuehl, 2013). However, PE has been found to differ from human translation as well as revision of human translated texts both in terms of the cognitive processes and the practical goals and processes (Krings, 2001; O’Brien, 2002). This likely leads to there being skills that are specific to PE.

In her proposal for PE course content, O’Brien (2002) adds to the general skills mentioned
above also certain a list of specific PE skills, including general knowledge of MT technology, terminology management skills, knowledge of pre-editing and controlled language, some programming skills, and text linguistic skills. O’Brien (2002) also notes the importance of a positive attitude toward MT, which is brought up also by many other writers (Rico and Torrejón, 2012; Doherty and Moorkens, 2013; Pym, 2013).

Rico and Torrejón (2012) discuss their view of necessary PE skills that are divided into three groups: core competences, linguistic skills, and instrumental competence. Core competences are “attitudinal or psycho-physiological competences”, which relate to dealing with subjectivity in PE specifications, client expectations and uncertainty, as well “strategic competence” for reaching informed decisions regarding PE alternatives. Linguistic skills relate to source and target language, communicative and textual skills as well as cultural, intercultural and subject area competence, and instrumental competence involves various technical skills related to understanding MT technologies, terminology management, corpora and controlled languages, as well as some basic programming skills.

Pym (2013) also discusses the general skills necessary for working with MT. He emphasizes the need to “learn to learn”, or learn how to pick up any new software quickly, as specific tools and skills related to them soon become outdated due to technological development. Related to this is the skill of evaluating the tools on offer. Another necessary skill set is learning to evaluate the MT suggestions (as well as translation memories) and the usability of segments, learning to make only necessary changes and to discard suggestions that require too many changes. Pym (2013) also suggests some specific revision skills involving detecting and correcting suprasentential errors (punctuation, cohesion) as well as stylistic revision, working as a part of a review team, and revising to a specified quality level.

Post-editing and MT have been included in at least some translation courses, often in the context of translation technology teaching (for example, Klifler, 2005, 2008; Austermeuhl, 2013; Fersoe et al., 2013; Kenny and Doherty, 2014; Doherty and Kenny, 2014). A commercial PE training course with assignments in various languages is also offered by the TAUS resource center 1. In translator training, MT and PE may also be used as part of larger projects: for example, Shuttleworth (2002) discusses a course involving a large translation project, where MT and PE were used together with translation memory and terminology tools. Torrejón and Rico (2002) describe a hands-on MT and PE exercise for a translation course. This exercise takes the form of a full translation assignment, starting from the definition of the assignment (client, text, schedule, purpose and style of the translation), followed by an error analysis of the MT output and pre-editing the ST to produce a new, improved MT version to be post-edited in accordance with PE guidelines.

Doherty et al. (2012) describe a course consisting of lectures and practical sessions focusing on statistical machine translation (see also Doherty and Moorkens, 2013). The topics forming the course content include introduction to MT history and concepts underlying the systems, MT evaluation, and the role of statistical MT and humans in the workflow (including pre- and post-processing of MT). Doherty and Moorkens (2013) also report on the student evaluations of the sessions, as well as themes arising from experiences during the course.

Kenny and Doherty (2014) also discuss some of the specific things translators should understand about statistical MT, where human translators fit in the workflow and what interventions that may be fruitful. Doherty and Kenny (2014) examine how these aims can be operationalized and included in translator training syllabus. The syllabus outlined in Doherty and Kenny (2014) covers both translation memories (TM) and MT, and includes topics such as basic concepts and implementation, evaluation of tools and MT output, pre- and post-processing, as well as professional issues like ethics, payment and collaboration.

---

1https://postedit.taus.net/post-edit/training-certification
Flanagan and Christensen (2014) describe a MT and PE course module consisting of two workshops introducing the translator students to MT and PE, and two PE assignments. The assignments involved PE according to PE guidelines where two different PE quality levels are defined (TAUS, 2010), and Flanagan and Christensen (2014) discuss how the translator students interpreted and followed these guidelines in their assignments.

To the best knowledge of the author, courses specifically focusing on PE have not been previously offered at Finnish universities, although PE exercises have been incorporated as parts of translation or translation technology courses. One example of such an exercise is discussed by Salmi and Koponen (2014). The exercise involved a for the students to translate a passage of English newspaper text into Finnish using two different MT systems, then compare the results and post-edit the version they considered easier to edit.

4 Course description

The course on post-editing was arranged at the University of Helsinki in the fall term 2014. It is an intermediate level course offered as part of the Translation Studies curriculum, and can be taken either by Bachelor’s level students as part of their minor, or by Master’s level students as a specialized course. The practical organization of the course consisted of seven two-hour lectures, five homework assignments and some in-class exercises, and a reflective essay written at the end of the course. The Moodle platform was used for distributing materials and collecting assignments.

The following subsections describe the students who took part in the course, the objectives of the course and the topics covered, and the practical exercises.

4.1 Students

The course was attended by 15 students, 13 of whom completed the final reflective essay. Translation Studies courses offered at the University of Helsinki are not limited to translation in any specific language pair or students of specific languages, but rather intended to support the studies of various language and translation-related majors. Therefore, the students’ study backgrounds included various language and translation subjects (English, German, Russian, Swedish and Nordic languages, Baltic languages). One of the students was from a more technical background, majoring in Language Technology. The language of teaching in the course was Finnish. The students were native speakers of Finnish, with the exception of one non-native speaker, also fluent in Finnish. All were fluent in English.

4.2 Course objectives and topics

The course aimed to introduce the students to the use of MT and PE in the translation industry, as well as research in the post-editing field. The course combined both theoretical background and practical exercises. The topics covered during the course include:

- Theory and history of MT and PE
- Practical use of MT and PE
- Controlled language and pre-editing for MT
- Post-editing without source text
- Post-editing process research
- Post-editing quality levels and guidelines
- MT quality evaluation and PE effort
• PE competences

In line with some of the recommendations and proposals discussed in Section 3, it was considered important to provide the students with knowledge of the theoretical principles of MT, as well as historical and practical context to the use of MT and PE. In particular, the outline can be compared to Doherty and Kenny (2014). The second topic, practical use of MT and PE, addressed also some of the professional issues brought up by Doherty and Kenny (2014), such as ethical questions and compensation in relation to the use of MT and PE. No lectures or practice on programming in general or building an MT system as such were included due to the limited course time and resources. On the theoretical side, we also considered it useful to provide the students with an overview of research concerning PE processes and competences, particularly the commonalities and differences compared to translation.

The theoretical material covered during the course was supplemented with five homework assignments involving post-editing machine translated texts and evaluating the quality of the machine translations, as well as some practical in-class exercises. The languages used in the exercises were mainly Finnish and English, with most post-editing done from English into Finnish. For some exercises, the students were offered options for using some other language pairs in the exercises.

4.3 Course assignments

The course assignments, as well as the texts, MT systems and other tools used are discussed below. Some general observations about the assignments and students’ comments are discussed in Section 4.5. As part of each assignment, the students were also asked to write a brief comment, which they could then later use as a starting point for their reflective essays.

Comparison of MT versions: The objective of the first assignment was to familiarize the students with different types of MT systems: rule-based MT and statistical MT. The theoretical background to these MT principles was discussed in the lecture preceding the assignment. To observe the differences in action, the students were given two MT versions of the flu treatment text, provided as MS Word documents. One MT version was produced using the statistical system (Google) and one by a rule-based system (Sunda). The students were asked to compare the two versions, examine their differences, and then post-edit the version they considered easier to edit. For this exercise, the students were instructed to produce a translation that could be published on the Internet for any reader looking for information on the flu, but no explicit PE guidelines were given.

Pre-editing: The objective of the second assignment was to introduce the concept of pre-editing. The students were given a brief overview of potential issues and a summary of commonly used English Controlled Language rules. They were asked to take at least five of the most problematic sentences from the text used in the previous assignment, try rewriting the sentences following the Controlled Language rules presented, and translate the rewritten versions with the MT systems used in the first assignment. The option of machine translating the sentences to some other target language was also given. As a further in-class exercise, the students were also able to use the terminology editor included in the Sunda tool in order to test the effect of user-specified terminology on the MT output.

PE without source text: The third assignment involved a scenario where the post-editor has no access to the source text. The students were introduced to research investigating the potential for such a scenario and its possible practical uses. For the assignment, they were given only the raw Finnish MT version of the dish washing liquid text, translated by the European Commission system. They were instructed to edit it so that it was grammatical and conveyed the meaning as they understood it; publication quality was not required. The text was again provided in a MS Word document. The students were also asked to track and report their own
time use in this assignment.

**Quality levels:** The objective of the fourth assignment was to introduce two different PE quality levels. The students were provided with the TAUS guidelines (TAUS, 2010), where two quality levels are defined. For the assignment, they were asked to post-edit two text passages, one according to the guidelines for “good enough” quality (semantically correct, no information added or omitted, no offensive or inappropriate content, correct spelling, corrections for stylistic reasons or fluency not required) and the other according to “human translation quality” (grammatically, syntactically and semantically correct, correct terminology, no information added or omitted, no offensive or inappropriate material, correct spelling and punctuation). For this assignment, the two press release passages were used. The students were divided into two groups and the text passages were alternated so that one group edited passage A according to the guidelines for “good enough quality” and passage B according to the guidelines for publication quality, and the text passages were reversed for the other group. This assignment was carried out using the Appraise tool, which also records PE time per sentence.

In class after the assignment, the students were able to compare their own PE time data to others in the class. They were also introduced to the use of edit distance metrics like HTER (Snover et al., 2006), which compare the number and type of changes made during post-editing. The students then examined their own edits using the metrics in Asiya Online Toolkit.

**MT quality evaluation:** The objective of the fifth assignment was to introduce the students to the topic of MT quality and evaluation, and to have them consider the distinction of fluency of language and adequacy of meaning in translation. They were provided with the LDC guidelines for evaluating fluency and adequacy (LDC, 2005), and asked to rank different MT versions of sentences either based on fluency or adequacy. For this assignment, the tourist phrases were used, and the evaluation was carried out using the Appraise tool. The students were offered different possible evaluation tasks in Finnish, Swedish, French, or German.

### 4.4 Evaluation of the practical organization of the course

Some issues related to planning and teaching the class can be noted. The course was planned to include many different types of practical exercises to provide the students with a wide view of different topics. On the other hand, the limited course time made it difficult to deal with many issues in depth, and some of the topics remained at a rather superficial level. The selection of texts and tools was complicated by the fact that the translator students’ main working language differed, and providing them all with materials specific to their own language pairs was not feasible within the scope of this course. The language pair English-Finnish was settled on for most exercises, as these were languages in which all had at least a good level of proficiency.

For texts to be edited during the course, we aimed to select relatively general interest texts that would not require particular expertise. One text consisted of 17 sentences (263 words) extracted from a longer informative text with instructions for treating the flu, intended for the general public. The second text consisted of 11 sentences (281 words) taken from a European Commission text on requirements for dish washing liquids. The third and fourth texts were extracts from English EU press releases. For the evaluation task, machine translated sentences were taken from the tourist phrasebook dataset described in Rautio and Koponen (2013). The set contains short, simple sentences involving asking directions, making purchases and similar, translated using different MT systems.

The selection of MT systems was limited by the availability of systems for Finnish. In the end, translations used in the PE exercises were produced with three systems: the statistical system developed by the European Commission, a rule-based system by Sunda Systems\(^2\), and

\(^2\)[http://www.sunda.fi/kaantaja.html]
Google Translate\(^3\). Texts translated with the European Commission system were included to represent perhaps the most likely intended scenario for PE: using an in-house system to translate specialized texts. The Sunda system is customizable through the use of user-specific terminology, and is reportedly used by some Finnish freelance translators and companies. Google was included mainly due to its familiarity and easy availability, although its usability for real-life PE scenarios is uncertain.

In terms of tools used, two of the PE tasks were given to the students in MS Word document format. One of the PE assignments and the evaluation task were carried out using the Appraise evaluation tool (Federmann, 2012). Although this tool is mainly intended for MT evaluation rather than PE, it provides basic functionality for post-editing sentences, and records PE time, for example. As an additional tool, a class exercise utilized the Asiya Online Toolkit (González et al., 2012), which is an online interface for using various automatic MT evaluation metrics.

The initial plan for the course had been to use a translation memory software for carrying out the assignments during the course. However, due to unexpected technical issues combined with the fact that surprisingly, most students indicated they had very little practical experience with the use of TM tools made their use not feasible during this course. This situation reflects previous observations (Doherty and Moorkens, 2013; Doherty and Kenny, 2014) about the importance of technical resources for teaching, and the challenges related to them.

### 4.5 Evaluation of learning outcomes

The objective of the PE course module described was to introduce the students to the use of MT and PE in the translation industry, as well as research in the post-editing field. The course aimed to provide the students with general knowledge of MT principles, the use of PE in practical scenarios and related professional issues, and related issues such as controlled language and pre-editing. Further, the goal of the course was to foster a positive attitude towards technology, together with the ability to critically evaluate the tools and processes, both of which have been found important (O’Brien, 2002; Pym, 2013; Kenny and Doherty, 2014). This section discusses these objectives in light of the class discussion as well as the reflective essays written by the students after the course. The reflective essays are discussed in more detail in Section 5.

Based on the first assignment where the different MT versions (rule-based and statistical) were compared, all students appeared to have grasped the basic principles of MT, and the differences between these two approaches. They were able to analyze the types of errors found in each version and reflect them against their understanding of the type of MT technology in question. In general the students found both versions surprisingly understandable, although not perfect. Some said there was “not much” to edit, that the MT provided a surprisingly good starting point, and that they felt post-editing had saved time and was not particularly challenging. Others, however, felt that they probably would have translated the text at least as fast.

The students were also able to assess the use of MT and PE critically. Most students doubted particularly the feasibility of PE without source text, based on their experience. Although they considered the text relatively understandable, they found it hard to trust the MT, and many pointed out that it might have been faster to translate the text using the source text. The students also did not consider pre-editing particularly beneficial. Although the application of some of the rules improved some sentences, or parts of sentences, many errors remained, and many students commented that it seemed rather futile to spend time pre-editing if one still had to do post-editing. They also felt that the introduction to controlled language in the preceding class had not been detailed enough.

The students also appeared to gain awareness of their own editing processes, and evaluate their own work. Most students commented that they found it somewhat difficult to differentiate

\(^3\)http://translate.google.com
between the PE quality levels, and to determine what changes were necessary for the “good enough” quality. Examining their own corrections, many commented that they likely were correcting “too much”. The in-class exercise examining the students’ own PE times and edits was overall very well received. The students commented that it was interesting and informative to compare the PE times and examine their own corrections after the fact. The students also commented that the evaluation task, where they rated MT quality focusing on language only or meaning only made them more aware of the issues.

Based on the course assignments and the students’ reflective essays, the course appears to have succeeded in introducing them at least to the basic principles of MT and PE. The students had a positive attitude towards the tools and workflows, but also evaluated them critically. As the course described here was a relatively short (7 weeks), many of the topics could only be discussed on a relatively superficial level. The PE assignments were also relatively short, and more extensive practice will naturally be required to make the students ready for PE work. Nevertheless, the students’ comments also show an awareness of their own editing processes, which will hopefully help them to further develop their skills.

5 Themes identified in the students’ reflective essays

At the end of the course, the students were asked to write a reflective essay on the central ideas and experiences during the course. They were instructed to write a 1000–1500 word essay discussing the theoretical background covered during the course and reflecting on their own experiences and observations from the assignments. As a starting point for the essay, they were asked to reflect on the following questions:

1. How did your understanding of the use of MT post-editing and related phenomena developed during the course?

2. What benefits and opportunities do you see in the use of MT and PE from the perspective of a translator, an organization requiring translation services, an individual MT user? What about problems or limitations?

3. How has the field developed and how do you believe it will develop in the future?

In addition, the students were instructed to raise other issues based on their readings and observations. They were also asked to comment on the course content, practical organization and potential improvements.

These reflective essays were analyzed for themes raised by the students. Similarly to the qualitative evaluation discussed in Doherty and Kenny (2014), extracts of the topics and observations discussed by each student were compared to determine whether multiple students brought up similar issues, and commonly occurring topics were grouped into larger themes. The themes related to the students’ expectations and attitudes toward MT and PE, observations on MT quality and usability, PE processes, PE levels and requirements, the students’ visions for the future of translation and PE, and training. The following subsections discuss these themes in more detail.

5.1 Expectations and attitudes

One clear theme commented on by the students related to their expectations and attitudes before the course, and some of the ways they changed. About half (7 out of 13) of the students stated that they had very little knowledge about MT and PE before the course. Most were at least somewhat familiar with free online MT systems, but as one student stated, the experience was mainly limited to laughing at examples of silly MT errors. One of the students had some prior theoretical and practical knowledge of MT, and another mentioned that the basic principles had
been covered in a prior course taken at another university. One student had worked in a company where MT was in use, although she had not used it in practice.

In terms of attitudes, most did not expect the quality of MT to be particularly high, and had not assumed it would be used, or usable, in professional contexts. Three of the students specifically described their own attitude toward MT as negative at the beginning of the course. Only one student mentioned having a generally positive attitude toward MT and PE. During the course, their perception of MT apparently changed, with most commenting that they saw more potential in MT after the practical experiences of the course.

Student attitudes was also one of the central themes identified by Doherty and Moorkens (2013), and in general attitude toward MT has been considered important in MT and PE related training. In this sense, the course appears to have succeeded in fostering a positive attitude, as most of the students commented on seeing MT in a more positive light after the course.

5.2 MT quality and trust

As noted in the previous subsection, most of the students did not have very high expectations of MT quality. In fact, most of the students (10 out of 13) explicitly stated that the quality of the machine translations handled during the course was better than they had expected. This could already be seen in the first assignment, where students commented that although neither version was “publication-ready”, containing some unintelligible parts, both were more understandable than they had expected. Their comments were even more positive concerning the texts translated using the European Commission system.

On the other hand, although positively surprised by the readability, the students found it hard to trust the MT. This point, which was raised in the essays by more than half of the students (8 out of 13), had also come up in class in connection with the PE assignment without source text. The students pondered whose responsibility the correctness (with regard to meaning) would be, and stated they would not feel secure taking responsibility for the final translation. The trust issue was not limited to this task, and more than one student discussed the slowing effect of constantly having to check the MT version against the source text, looking for potential errors.

It should be noted that although the students appeared to be even somewhat impressed with the quality of the MT used, all the texts still contained various issues, as MT quality into Finnish remains far from fluent. Similarly to the theme of quality thresholds discussed by the students in Doherty and Kenny (2014), the student of this course also identified situations where the MT quality was likely too poor for any intervention to be fruitful. The MT texts, however, appear to be at the level deemed most usable for teaching purposes by Kliffer (2005): not bad enough to contain such frequent low-level errors that PE would be completely pointless, but not such high quality that there would be hardly anything to correct.

5.3 PE process and time

Half of the students (7 out of 13) discussed the unfamiliarity of the PE work compared to their prior experience with translation, and stated that it took time to get used to the idea of working with a “raw” translation. The difficulties mentioned involved, for example, getting used to the MT differing so greatly from a raw version they might create themselves, and having to deal with two sources, the source text and the MT. Some also commented that the sentence-by-sentence way of working imposed by the Appraise tool was difficult, while others found it intuitive. One even stated he started to quite like using the tool.

As could be seen in the third assignment, where the students tracked their own time, their speed varied greatly. Of the 13 who completed the assignment, 4 had taken less than 30 minutes, 5 had taken between 30 and 60 minutes, and 4 students had taken more than 60 minutes to edit the relatively short passage. Some of the differences in PE speed may of course be related to
the students’ overall language skills, as the language pair English-Finnish did not represent the main working languages for all of the students. A more detailed analysis of the relationship between language skills, edits and PE times was not within the scope of this paper, but might be informative in the future.

Based on their class comments and the reflective essays, they also differed in how they viewed their time use. Some stated that once they started to feel more comfortable, they felt they were saving time compared to translation from scratch. One student estimated that PE probably took only about one third of the time that translating the same text passage would have. On the other hand, not every student felt the PE was making them faster. One of the students even expressed some frustration and feeling she was still working too slow and making unnecessary changes.

The students’ observations can be compared to the findings of the survey reported in Salmi and Koponen (2014), where most stated they either would have translated the same text faster or that there was no difference in time. Only 10 of the 49 students answering that survey felt PE made them faster. In studies where speed in PE versus translation has been explicitly compared, García (2010, 2011) have indeed found that for translation students, PE was only marginally faster. These experiments, however, were conducted with students who had no training in PE, so partly this may be due to unfamiliarity.

Interestingly, one student explicitly connected her PE speed to the amount of experience, noting that the PE probably speeded up her work specifically because she is not yet a particularly experienced translator. She wrote in her essay that if her translation processes had been more established, she might have found it more difficult to work with the MT. No detailed analysis was performed on the students’ edits and it is therefore difficult to say how proficient this particular student was, but her self-assessment reflects the observation by García (2011) that the “poor” students benefited more from MT, and that one possible explanation was that they had not yet formed effective translation skills. On the other hand, as García (2011) points out, these students could also have some other characteristics that make them particularly suited for PE.

5.4 Adjusting to PE quality requirements

As with MT and PE in general, the students indicated that they had relatively little knowledge of the different use scenarios and quality levels for PE. Particularly the idea of “good enough” quality was new to them, and most (9 out of 13) explicitly commented that they had a hard time adjusting to a task where such quality was required. Thinking of the translation as a text that needed to be simply sufficient to convey the meaning needed some adjustment. However, most indicated that after the course, they now had a better understanding of differing situations where a lesser quality of language might suffice. Most also explicitly mentioned that they considered it important to learn to adapt to different requirements.

Most of the students (7) who raised this theme also explicitly connected it to their prior translator training, where the need for “perfect” language is generally emphasized. They found it difficult to accept a translation that was “less than perfect”. The students appeared to react to this in two ways. Some wanted to work on the translation until they were happy with the final version, but recognized that they were probably “editing too much”, which was also connected to their observations on time use. Referring specifically to the first PE assignment, one student stated her final version should properly be called “inspired by the MT” rather than “post-edited”. Other students decided to try to manage only with minimal corrections, and not try to make the text their “own”, but expressed concerns about how the often unidiomatic MT affected their final translations.

The concerns about quality of the final translation are reflected, for example, in the survey
reported by Salmi and Koponen (2014), where many of the students expressed similar issues. Their apparent desire to edit the final translation to be as perfect as possible, even if it meant rewriting, however, contrasts with the findings of Depraetere (2010). Based on the analysis of the corrections made by students in a PE project, Depraetere (2010) reports that the students “did not feel the urge to rewrite it”, and did not rephrase the text if the meaning was clear. The observations regarding difficulty of adjusting to the quality levels can also be compared to Flanagan and Christensen (2014), whose students also discussed making potentially “unnecessary” stylistic changes, even when the guidelines called for “good enough” quality. Similarly to those students, our students struggled with adhering to the guidelines related to style. As noted by Flanagan and Christensen (2014), it may also be that the guidelines themselves are not entirely clear.

5.5 Future visions

Since the essay instructions had specifically asked the students to reflect on potential future visions, all of the students did discuss views regarding MT and PE both as a general practice and in their future careers. Of the 13 students, 11 students explicitly stated they believed the use of MT and PE would increase. In contrast with their rather low initial expectations for MT, and even self-professed negative opinions before the course, most seemed to have a much more positive outlook for such future. One student who described her view as very negative before the course even stated she had completely changed her mind on MT. Over half (8 out 13) of the students explicitly stated that they viewed MT as a tool, not a threat for them as translators.

Nearly all, however, also discussed the limitations of PE, particularly of any scenario involving PE without source text. The essays also reflected a clear understanding that MT and PE are not feasible in every situations, which is an important point related to the need of helping students become also critical users of technology, as argued by Kenny and Doherty (2014).

Some of the students also explicitly commented on how they saw PE as a potential career prospect. Five of the students stated that they could well see themselves doing PE work in the future. Two commented that they would not be interested in PE work; one specified that she was mainly interested in literary translation, where MT is likely to be of little use. The other six students did not comment on PE in terms of their own future plans.

5.6 PE training and potential improvements to the course

Nearly all of the students also explicitly commented on MT and PE related training. In general, they considered the topics and competences covered important, even essential for translator training. Even the students who stated that they did not consider PE an attractive career option for them personally felt the course would overall benefit translator trainees. One of them noted that it was good to know what the work is like, so she can better decide whether to accept potential jobs offered in the future.

As this course was the first time a specific PE course was taught, it was very important to hear the students’ views on how the course contents and practical organization worked. Most stated they were overall very satisfied with the course and found the topics interesting. The chance to examine one’s own PE time data and edits was mentioned as a particularly interesting and useful exercise. Most commonly mentioned improvement was that future courses should include options for more varied language pairs. Some students also hoped that more concrete examples would have been used in introducing topics like controlled language and PE levels, and that there would have been more detailed discussions of specific examples from the edited texts. With regard to specific topics, many found pre-editing least interesting and useful. Some students did not find PE without source text particularly useful, whereas one hoped even more class time had been spent on that topic.
6 Conclusion and future teaching

In this paper, we have described experiences from a translator training course focusing on MT and post-editing. The course was taught for the first time during fall term 2014 at the University of Helsinki, Finland. After a brief overview of the context of MT and PE use in the country, and an overview of related work on teaching MT and PE, we have presented a description of the course contents, practical organization and assignments intended to introduce the students to various theoretical topics and practical issues in the field.

Common themes arising from reflective essays written by the students after the course were also identified and discussed in combination with observations and comments from the assignments and class discussions. The central themes identified involved students’ expectations and attitudes, MT quality and trust, the PE process and time use, adjusting to PE quality requirements, future visions regarding PE, and PE training. A particularly interesting result was seen in the students’ attitudes, in that their generally low expectations and even negative attitudes toward MT before the course appeared to change to a much more positive view. However, they also expressed rather realistic views on the potential and limitations of MT.

The materials and tools selected for the course appeared mostly suitable for the purpose. Although MT quality into Finnish still lags behind many others, the systems used were able to produce translations of the source texts selected with high enough quality to make PE feasible. Two of the MT systems used were known to be used in practice by at least some translators working in the field, which was considered important for a realistic view of the situation. When selecting the MT systems to be used, Google was included due to its familiarity and availability to the students. Interestingly, the survey by Gaspari et al. (2015) has found that the vast majority of their respondents who used MT, used free online MT systems like Google or Bing. Although it is difficult to say whether that includes translators working into Finnish, there is also anecdotal evidence pointing to its use in this country.

Some practical and technical challenges were also observed. The most important challenges relate to the fact that the course is offered to students in all language and translation subjects, which makes it difficult to provide materials for all the language pairs they may be working in. Another issue involved technical challenges, and the inability to use translation memory tools for the assignments.

The course is being taught again in Fall term 2015. Based on the experiences of the course in 2014, we have implemented PE exercises in more varied language pairs. To better reflect the practical scenario, most of the assignments are also carried out using a TM system. The exercise where the students were able to examine their own edits and PE times appeared particularly useful, and we plan to incorporate this aspect already in the earlier assignments. We hope that this may help the students observe their own processes more closely over several assignments. Together with class discussions, they may be then able to try different approaches in the following assignments. The ordering of the assignments will also be reconsidered. In particular, the adequacy versus fluency evaluation task might be more fruitful in the beginning of the course, as it could assist the introduction of PE requirements and “good enough” quality, which the students found particularly hard to accept.

Acknowledgements

The author wishes to thank all the students for participation in the course, and the European Commission, Sunda Systems Oy, and professor Lauri Carlson for offering tools and materials for use in the course.
References


Quality Assessment of Post-Edited versus Translated Wildlife Documentary Films: a Three-Level Approach

Carla Ortiz-Boix carla.ortiz@uab.cat
Anna Matamala anna.matamala@uab.cat
Department of Translation and Interpretation & East Asian Studies, Universitat Autònoma de Barcelona, Bellaterra, 08193, Spain

Abstract
This article presents the results of a study designed to evaluate the quality of post-edited wildlife documentary films (in comparison to translated) which are delivered using voice-over and off-screen dubbing. The study proposes a quality assessment at three levels: experts’ assessment, dubbing studio’s assessment and end-users’ assessment. The main contribution of this quality assessment proposal is the inclusion of end-users in the process of assessing the quality of post-edited and translated audiovisual texts. Results show that there is no meaningful difference between the quality of post-edited and translated wildlife documentary films, although translations perform better in certain aspects.

1. Acknowledgements
This article is part of the ALST project (reference FFI-2012-31024) led by Anna Matamala and funded by the Spanish "Ministerio de Economía y Competitividad" and it is also partially funded by research group TransMedia Catalonia (2014SGR27) and the grant FI-DGR2013 awarded by AGAUR. This article is part of the research carried out by Carla Ortiz-Boix under the supervision of Dr. Anna Matamala within the PhD in Translation and Intercultural Studies from the Department of Translation, Interpretation & East Asian Studies at Universitat Autònoma de Barcelona.

2. Introduction
Quality and quality assessment (QA) have been a central issue in Translation Studies since the beginning of the discipline. Many studies have been carried out in that regard (e.g. Nida, 1964; Reiss et al, 1984; Gambier, 1998; Hansen, 2008; Melby et al, 2014), approaching both quality and QA differently depending on the translation theory (House, 2006). Studies on machine translation (MT) and post-editing (PE) have also addressed quality and QA by developing models and measures to evaluate the quality of the text types (technical and general) in which MT and PE is most frequently applied. Although recent studies (Melero et al, 2006; Bywood et al, 2012; Etchegoyhen et al, 2014; Fernández et al, 2013; Ortiz-Boix and Matamala, forthcoming) have proved that including MT and MT plus PE into the workflow of some audiovisual translation (AVT) modalities, mostly subtitling, would positively impact productivity, research into quality and QA of both MT and PE in AVT is still much needed.
This article presents an experiment in which the quality of post-edited wildlife documentary excerpts delivered through voice-over (VO) and off-screen dubbing (OD) has been assessed in comparison to the quality of translations of the same wildlife documentary excerpts. This experiment has been carried out because, after research by Ortiz-Boix and Matamala (forthcoming) demonstrated that applying post-editing instead of translation in these transfer modes could be feasible in terms of effort involved, it is yet to be known how this would impact on quality. Our QA proposal takes into account the specificities of the two audiovisual transfer modes involved (VO and OD) and includes a new aspect that has been usually left aside: the involvement of end-users. It also includes a brief quality assessment by the dubbing professionals that recorded the translated and post-edited versions that were used afterwards in the user reception test.

In order to contextualize our experiment, Section 3 briefly describes the two audiovisual transfer modes under analysis, and summarizes how post-editing QA, and QA in AVT have been approached so far. Section 4 describes the methodological aspects of our QA test. In Section 5, results are presented, and conclusions and further research are discussed in Section 6.

3. Previous Work

This section defines VO and OD, highlighting the specificities of these AVT modalities (3.1). It then summarizes previous work on post-editing QA, with an emphasis on audiovisual translation that has inspired the study (3.2).

3.1. Voice-Over and Off-Screen Dubbing

VO is the AVT transfer mode that revoices an audiovisual text in another target language on top of the source language voice, so that both voices are heard simultaneously (Franco et al, 2010). In countries such as Spain, VO is the transfer mode frequently used in factual programs, e.g. documentary films, as it is said to help reproduce the feeling of reality, truth and authenticity given by the original audiovisual product (Franco et al, 2010). In Eastern Europe, however, VO can also be found in fictional TV programs.

OD is the transfer mode that revoices off-screen narrations substituting the original voice with a version in the target language (Franco et al, 2010). In other words, when OD is applied, only the target language version is heard, not the original one. OD is used in factual programs and usually combined with VO (OD for off-screen narrators, VO for on-screen interviews).

Some of the main features of these transfer modes are the following:

1) Both VO and OD present synchronization constraints. In VO three types of synchrony are observed: kinetic synchrony – the translated text matches the body movements seen on screen–, action synchrony – the translated text matches the actions seen on screen–, and voice-over isochrony – the translated message fits between the beginning and the end of the original speech, leaving some time after the original voice starts and before it ends where only the original can be heard. OD is only endowed with kinetic and action synchronies, as the original voices are not heard in this transfer mode (Orero, 2006; Franco et al, 2010).

2) Different language registers can coexist in audiovisual productions where VO and OD are used: whilst VO is generally used for semi-spontaneous or spontaneous interviews, OD is usually applied to narrators with a planned discourse (Matamala, 2009; Franco et al., 2010). If the original product contains oral features such as fluffs, hesitations and grammatical mistakes, the target language version does not generally reproduce them (Matamala, 2009). In other words, the translation is generally an edited version of the original.
VO and OD are often used to revoice wildlife documentary films from English into Spanish, the object of our research. This type of non-fictional genre usually includes many terms that might pose additional challenges to the translators (Matamala, 2009). It is also often the case that the source text contains linguistic errors and inconsistencies (Franco et al, 2010), and that a quality written script is not available (Ortiz-Boix, forthcoming). However, translators are expected to deliver a quality written script in the target language so that the recording by voice talents in a dubbing studio can begin.

3.2. Post-Editing Quality Assessment

Although research on QA of post-edited text has increased, it is still rather limited. Fiederer and O’Brien (2009), Plitt and Masselot (2010), Carl et al (2011), García (2011), Guerberof (2009, 2012), Melby et al (2014) and Mariana (2014) have dealt with quality in post-editing, to a greater or lesser extent. Up until now, QA has been based mostly on what has been has termed in the QTLauchPad project (Lommel et al, 2014) as either holistic approaches – which assess the quality of the text as a whole – or analytic approaches – which assess the quality by analysing the text in detail according to different sets of specifications. A combination of both can also be found.

**Holistic approaches:** Plitt and Masselot (2010) used the Autodesk translation QA team to assess randomly selected samples of translated and post-edited text using two labels (“average” or “good”), depending on whether they considered the text was fit for publishing. In Carl et al (2011), raters ranked the quality of a list of sentences, either translated or post-edited. Fiederer and O’Brien (2009) also assessed the quality of sentences – three translated and three post-edited versions of 30 sentences – according to clarity, accuracy and style on a 4-point scale. Raters were also asked to indicate their favorite option out of the six proposals for each source sentence.

**Analytic approaches:** In García (2011), a rater assessed the quality of a 500-word text by using the Australian National Accreditation Authority for Translators and Interpreter’s (NAATI) guidelines. In Guerberof (2009, 2012), three raters blindly assessed translated segments, post-edited segments and segments previously extracted from a translation memory by using the LISA QA model.

**Mixed approaches:** Melby et al (2014), Mariana (2014) and Lommel et al (2014) develop and implement the Multidimensional Quality Metrics (MQM) in their analysis. The model provides a framework for defining metrics and scores that can be used to assess the quality of human translated, post-edited or machine translated texts. It sets error categories, otherwise called issue types, which assess different aspects of quality and problems. MQM is partly based on the translation specifications (Melby, 2014) that define expectations for a particular type of translation; MQM is organized in a hierarchical tree that can include all the necessary issue types for a given text type and a given set of specifications.

In the specific field of audiovisual translation, post-editing quality assessment research is still more limited: EU-financed project SUMAT (Etchegoyhen et al, 2014) evaluated the quality of the machine translation output via professional subtitlers who assigned a score to each subtitle. They were asked for general feedback on their experience while post-editing as well as on their perceived quality of the output. Aziz et al (2012) assessed the quality of the machine translated subtitles by post-editing them using the PET tool. The post-edited subtitles were afterwards assessed against translated subtitles using BLEU and TER automatic measures, suggesting there is no meaningful difference in terms of quality between them.
4. Methodology

Our experiment involved one language pair (English into Spanish), and aimed to assess the quality of post-edited wildlife documentaries compared to the quality of human translations. It is built upon the hypothesis that there is no meaningful difference between the quality of post-editing and the quality of translation of wildlife documentaries in English delivered through VO and OD in Spanish.

The experiment included a three-level quality assessment: (1) quality assessment by experts, with a mixed approach (holistic and analytic); (2) quality assessment by the dubbing studio where the translations and post-editings were recorded, and (3) quality assessment by end-users, who watched both post-edited and translated audiovisual excerpts. The inclusion of end-users in the assessment has been inspired by functionalist approaches to translation and by recent user reception studies in AVT. In the case of wildlife documentaries, we wanted to assess whether both post-edited and translated documentaries fulfilled their function to the same extent, that of informing and entertaining the audience.

4.1. Participants

Participants taking part on the first level assessment were six lecturers of MAs on audiovisual translation in universities in Spain who are experts on VO and currently work or have recently worked as professional voice-over translators. The experts’ profiles are comparable: all of them have a BA in Translation Studies except for one, who has a BA in German Studies. Furthermore, five of them have either a PhD in Translation or have attended PhD courses on the same field. Previous experience varies among participants: when the experiment was carried out experts 1, 3, and 5 had worked as audiovisual translators between 10 and 16 years and taught for 11, 8, and 5 years respectively, while participants 2, 4, and 6 had between 5 and 8 years of experience as audiovisual translators and taught for the last 4 or 5 years. The number of experts used to rate the documents is higher than in previous studies on QA and post-editing (Guerberof, 2009; García, 2011; or De Sutter et al, 2012).

For the second level, only one dubbing studio was used, as only one study was needed to record the materials. Two voice talents, a dubbing director and a sound technician were present during the recording session.

In the third level, 56 users with different educational backgrounds took part in the experiment (28 male, 28 female, 23-65 years old, mean age: 39.15). All participants were native speakers of Spanish and 46.43% of the participants were highly proficient in English. Watching habits related to wildlife documentaries do not vary much among participants (96.43% watch a maximum of 3 documentaries on TV every month), but preferences in terms of the audiovisual transfer mode to be used in wildlife documentaries differ: 30.46% prefer subtitling, 44.64% prefer dubbing, and 25% prefer VO. These preferences are correlated with age: participants under 40 prefer subtitled documentaries (50%), whilst participants over 40 prefer voiced-over documentaries (46.3%).

4.2. Materials

The materials used for the first level were 6 translations and post-editings of two self-contained excerpts of a 7-minute wildlife documentary film titled Must Watch: a Lioness Adopts a Baby Antelope that is currently available on Youtube as an independent video (http://www.youtube.com/watch?v=mZw-1BFHFM). It is part of the episode Odd Couples from the series Unlikely Animal Friends by National Geographic broadcast in 2009. Short excerpts were chosen for practical reasons, despite being aware that this could impact on
evaluative measures of enjoyment and interest. Additionally, excerpts of a wildlife documentary were chosen since documentaries follow structured conventions and have specific features in terms of terminology (Matamala, 2009). The translations and post-editings (24 in total) were produced by 12 students of an MA on AVT that had had a specific course on VO but no, or almost none, previous experience on post-editing. Hence, they were instructed to correct all the errors and adjust, only if necessary, the text according to the specific constrains of documentary translation. Participants worked in a laboratory environment that recreated current working conditions: they used a .doc document and they were allowed to use any available resources (internet, dictionaries, etc.) To perform both tasks, students were given a maximum of 4 hours, although almost none of them used the entirety of the given time. The audiovisual excerpts were similar in terms of length (first excerpt: 101 seconds, 283 words; second excerpt: 112 seconds, 287 words) and content, and the translations and post-editings contained between 218 and 295 words. They were machine translated through Google Translate, the best free online MT engine to be used to machine translate wildlife documentary scripts according to Ortiz-Boix (forthcoming).

For the second level, the best post-editing and the best translation of each excerpt was selected, according to the results of the first-level quality assessment. The recordings of these excerpts were used for the third-level assessment.

4.3. Test Development

Level 1: Experts’ Assessment. Participants carried out the experiment from their usual place of work. They were given detailed instructions on how to assess the 24 documents without knowing which of them were translated or post-edited. They were given 20 days to perform the whole assessment. The experiment was divided into three evaluation rounds:

a) In round 1, raters were instructed to read each document and grade it according to their first impression on a 7-point scale (completely unsatisfactory-deficient-fail-pass-good-very good-excellent). They were just given one day for this task, and the order of the documents was randomized across participants.

b) In round 2, raters were asked to correct the documents following a specific evaluation matrix (see section 4.4.), and grade them after the correction on a 7-point scale. Afterwards, they had to answer an online questionnaire (see section 4.5.).

c) In round 3, a final mark between 0 and 10, following Spain’s traditional marking system, was requested.

There was also a final task in which raters had to guess whether the assessed document was translated or post-edited (post-editing/translation identification task).

Level 2: Dubbing Studio Assessment. The scripts and videos were sent to the dubbing studio and a professional recording was requested from them. They were instructed to follow standard procedures. A researcher took observational notes and gathered quantitative and qualitative data on the changes made during the recording session by the dubbing director.

Level 3: End-Users’ Assessment. Quality was understood to be based on end-user reception and, following Gambier’s proposal (2009), three aspects were assessed: understanding, enjoyment, and preferences (or response, reaction and repercussion in Gambier’s terms). Participants were invited to a lab environment that recreated the conditions in which documentaries can be watched: they sat in an armchair and watched the documentary excerpts in a 32’ flat screen. Taking into account ethical procedures approved by Universitat Autònoma de Barcelona’s ethical committee, participants were administered a pre-task questionnaire (see section 4.6.). They were then shown two of the excerpts without
knowing whether they were watching a translated or post-edited excerpt. After each viewing, a questionnaire was administered to them to test their comprehension and enjoyment, as well as their preferences (see section 4.6).

4.4. Evaluation Matrix (Level 1)

The evaluation matrix applied in the first level is based on MQM because it can be used for both translations and post-editings, and it also allows to select and add only the relevant categories for our text type. Although MQM offers the possibility to include over one hundred issue types, only five categories and eleven subcategories of issue types were selected, as shown on Table 1.

<table>
<thead>
<tr>
<th>Issue types categories</th>
<th>Issue types subcategories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adequacy</td>
<td>- Wrong Translation</td>
</tr>
<tr>
<td></td>
<td>- Omission</td>
</tr>
<tr>
<td></td>
<td>- Addition</td>
</tr>
<tr>
<td></td>
<td>- Non-translated words</td>
</tr>
<tr>
<td>Fluency</td>
<td>- Register</td>
</tr>
<tr>
<td></td>
<td>- Style</td>
</tr>
<tr>
<td></td>
<td>- Inconsistencies</td>
</tr>
<tr>
<td></td>
<td>- Spelling</td>
</tr>
<tr>
<td></td>
<td>- Typography</td>
</tr>
<tr>
<td></td>
<td>- Grammar</td>
</tr>
<tr>
<td></td>
<td>- Others</td>
</tr>
<tr>
<td>Variety</td>
<td></td>
</tr>
<tr>
<td>Voice-over/off-screen</td>
<td>- Spotting</td>
</tr>
<tr>
<td>dubbing specificities</td>
<td>- Action and kinetic synchronies</td>
</tr>
<tr>
<td></td>
<td>- Phonetic transcriptions</td>
</tr>
<tr>
<td></td>
<td>- VO Isochrony</td>
</tr>
<tr>
<td>Design/Layout</td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Evaluation matrix: error typology

The selection was based on previous research on errors produced by MT engines in general texts (Avramidis et al, 2012) and wildlife documentary films (Ortiz-Boix, forthcoming), as well as in post-editings (Guerberof, 2009). As MQM does not contain a domain specific issue type for audiovisual translated texts, a new category was added: VO/DO specificities. It includes the issue types subcategories spotting, action and kinetic synchrony, voice-over isochrony, and incorporation of phonetic transcriptions. Raters were trained on how to apply the evaluation matrix.

4.5. Questionnaire design (Level 1)

The questionnaire in level 1 aimed to gather the agreement of the raters with eight statements assessing fluency, grammar, spelling, vocabulary, terminological coherence, voice-over specifications, satisfaction, and success in terms of purpose, using a 7-point Likert scale:

- In general, the text was fluent.
In general, the translation was grammatically correct.
In general, there were no spelling issues.
In general, the vocabulary was appropriate.
In general, the terminology was coherent throughout the text.
In general, the translation met the VO and DO specificities.
In general, the final result was satisfactory; aka the translation met its purpose.
In general, the translation could be sent to the dubbing studio to be recorded.

4.6. Questionnaire design (Level 3)

The pre-task questionnaire included five open questions on demographic information (sex, age, highest level of studies achieved, mother tongue, and other spoken languages) as well as seven questions on audiovisual habits.

The post-task questionnaire included seven questions on enjoyment. Participants had to report their level of agreement on a 7-point Likert scale on the following statements:

- I have followed the excerpt actively.
- I have paid more attention to the excerpt than to my own thoughts.
- Hearing the Spanish voice on top of the original English version bothered me.
- I have enjoyed watching the excerpt.

They also had to answer the following questions on a 7-point Likert scale:

- Was the excerpt interesting?
- Will you look for more information regarding the couple presented on the documentary?
- Would you like to watch the whole documentary film?

They were also asked 3 questions on perceived quality and comprehension, again on a 7-point Likert scale:

- The Spanish narration was completely understandable.
- There were expressive problems in the Spanish narration.
- There were mistakes in the Spanish narration.

Five additional open questions per excerpt were used to test comprehension. Finally, participants were asked which excerpt they preferred. A pilot test was run to validate the questionnaire, which was inspired by Gambier (2009).

4.7. Data and Methods

The following data were obtained:

Level 1 (experts):
1) 144 documents with corrections (6x24) according to the MQM-based evaluation matrix.
2) The grades for each document in the three scoring rounds.
3) 144 completed questionnaires (6x24 documents) reporting on the participants' views after correcting each document.
4) The results of the post-editing/translation identification task.

Level 2 (dubbing studio):
5) 4 documents with corrections (1x4) made by the dubbing director and their corresponding recordings.
6) Observational data gathered during the recording session.

Level 3 (end-users):
7) 56 completed questionnaires on demographic aspects and audiovisual habits.
8) 112 completed questionnaire responses (14x4) on user enjoyment, comprehension and preferences. In order to analyse the comprehension questionnaire, wrong answers were given 0 points, partially correct answers were assigned 0.5 points and correct answers, 1 point.

All data were analysed using the statistical system R-3.1.2, developed by John Chambers and colleagues at Bell Laboratories. In this study, data was analysed according to descriptive statistics.

5. Discussion of Results

Results are presented according to the three levels of assessment. More attention is devoted to levels 2 and 3, as a more detailed analysis of the first level is already presented in Ortiz-Boix and Matamala (forthcoming).

5.1. Quality Assessment by experts

The quality of both translations and post-editings was rather low and no meaningful differences between post-editings and translations in terms of quality were found, as the difference between the scores for each of the tasks were low. Results are discussed in two different sub-sections: in the holistic approach, the scores given in the evaluation rounds, the questionnaire replies and the identification task results are analysed. The analytic approach discusses the results of the corrections performed by the raters.

5.1.1 Holistic Approach

Results of round 1 indicate that experts evaluate better translations than post-editings after reading the documents for the first time: while 45 out of 72 (62.5%) translations were evaluated from "pass" to "excellent", only 37 out of 72 post-editings (51.39%) were evaluated within this range. However, when documents are rated again after a thorough correction (round 2), the difference between post-editings and translations was not significant. The table below shows the pass marks for each round:

<table>
<thead>
<tr>
<th></th>
<th>Passes for Round 1</th>
<th>Passes for Round 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translations</td>
<td>45</td>
<td>41</td>
</tr>
<tr>
<td>Post Editings</td>
<td>37</td>
<td>38</td>
</tr>
<tr>
<td>Total Possible</td>
<td>72</td>
<td>72</td>
</tr>
</tbody>
</table>

Table 2. Pass marks for round and task

---

1 See Ortiz-Boix and Matamala (forthcoming) for further information on the results of this level.
translations diminishes. In this case, 41 out of 72 translation (56.94%) and 38 out of 72 post-editings (52.78%) are given between a "pass" and an "excellent". Despite these slight differences, the median grade in both rounds is a “pass” for both translations and post-editings.

Results for round 3, in which the Spanish traditional marking system was used (from 0 to 10, 5 being a “pass”), show again a very small difference: the mean grade for translations is 5.44 versus 5.35 for post-editings. This mark correlates perfectly with grades obtained in rounds 1 and 2.

As for the questionnaire replies, results indicate that post-editings are given higher grades in four of the issue types – grammar, terminological coherence, satisfaction, and success in terms of purpose– and the exact same grade in the case of VO specificities. Translations are considered better in fluency, vocabulary appropriateness, and spelling. However, no relevant differences are found in any case.

Concerning the final identification task, experts correctly categorized 42 translations out of 72 (58.33%) and only 22 post-editings (30.56%). They categorized wrongly 14 translations (19.44%) and 27 post-editings (37.5%), and could not decide whether the document was a translation or a post-editing in the case of 16 translations (22.22%) and 23 post-editings (31.94%). Results indicate that post-editings are more difficult to identify than translations, as the great majority of them are either misidentified or not recognized as such. If the quality of post-editings were generally worse, a better identification would be expected, which leads us to suggest that the quality of both translations and post-editings is comparable.

![Figure 1. Task identification](image)

### 5.1.2. Analytic Approach: Correction

Translations present a lower number of corrections (mean: 12.861 per document) than post-editings (17.957), although the mean difference in a text is five corrections and it is not meaningful. It is interesting to highlight that experts did not correct any errors regarding synchrony and did a higher number of corrections for post-editings in all issue types but three: omission, addition, and spelling (see Ortiz-Boix and Matamala forthcoming for further details). The issue types with more errors, both in post-editing and translation, were wrong translation, style, typography, and grammar. Given the small differences, results seem to
prove that the quality of both translations and post-editings in our experiment was similar, although the type of errors found in either post-editings or translations differ.

5.2. Quality Assessment by the Dubbing Studio

During the recording session it was observed that changes made in the translation and post-editing scripts were only related to aspects directly linked to the voicing of the documentaries.

In the first excerpt, a similar number of changes were made: six in the post-editing excerpt, five in the translated excerpt. Changes referred to synchronization aspects (3 in the translated version and 2 in the post-edited one), phonetics (2 and 1 respectively), and stylistic repetitions (0 and 3 respectively); the experts in level 1 had surprisingly not corrected issues related to synchronization. In the second excerpt, 4 changes were made in the translated version (1 on phonetics and 3 on synchronization). As for the post-editing, the dubbing director pointed out that the synchronization was not good and that a re-translation was needed. However, it was decided to record it as it was and test whether audiences would react negatively.

Although no quantitative differences were observed, data show that the translation, at least in the second excerpt, was qualitatively better than the post-edited script.

5.3. Quality Assessment by Users

Data were analysed taking into account all participants but a more specific analysis divided participants into two age groups (group A: <40, group B: >40) as differences in terms of preferences for subtitling or VO were observed in the demographic questionnaire. Results are presented in terms of enjoyment and preferences (see section 5.3.1.) and understanding (see section 5.3.2.).

5.3.1. End-Users Enjoyment and Preferences

Results indicate that, regardless of the excerpt, version, and age group, users mostly agree with the fact that they followed the excerpt actively (median for all conditions/groups/excerpts = “strongly agree”) and focused on what they were watching on screen (all medians are “strongly agree”, except for post-editing of excerpt 1= “moderately agree”). Hearing the Spanish voice on top of the original English version did not bother any of the participants in any of the conditions or excerpts (median = “strongly disagree” with the statement “Hearing the Spanish voice on top of the original English version bothered me”), although percentages show a difference between age groups: older viewers (96.43%) are not bothered at all by the Spanish voice on top of the original English voice (“strongly disagree” with the statement), whilst the percentage in younger viewers drops (57.14%). This percentage, though, is distributed evenly across both versions, showing that it is the transfer mode (VO) and not the translation system (translation/post-editing) that impacts on them. This also correlates with the preferences stated by younger audiences in the pre-task questionnaire.
I have followed the excerpt actively

<table>
<thead>
<tr>
<th></th>
<th>Excerpt 1</th>
<th>Excerpt 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Translation</td>
<td>Post-editing</td>
</tr>
<tr>
<td>I have followed the excerpt actively</td>
<td>Strongly agree</td>
<td>Strongly agree</td>
</tr>
<tr>
<td>I have paid more attention to the excerpt than to my own thoughts</td>
<td>Strongly agree</td>
<td>Moderately agree</td>
</tr>
<tr>
<td>Hearing the Spanish voice on top of the original English version bothered me</td>
<td>Strongly disagree</td>
<td>Strongly disagree</td>
</tr>
<tr>
<td>I have enjoyed watching the excerpt</td>
<td>Strongly agree</td>
<td>Moderately agree</td>
</tr>
</tbody>
</table>

Table 3. Agreement level on enjoyment (medians)

When asked to express their level of agreement or disagreement with the statement “I have enjoyed watching the excerpt”, users grade the translated version higher than the post-editing one (translation median= “strongly agree”, post-editing median: “moderately agree”). Although there are slight differences depending on the excerpt: in excerpt 1, 57.14% of the participants strongly agree with the statement whilst the percentage drops to 32.14% in the post-editing, being the median “strongly agree” for the translation and “moderately agree” for the post-editing. In excerpt 2, differences in enjoyment are higher: 85.71% of the users who watched the post-edited version strongly or moderately agree with the statement, in contrast with 57.14% of the users of the translated version. The median for the post-editing is “strongly agree” and for the translation it is “moderately agree”. Slight differences are observed between age groups, since overall the younger group “moderately agrees” with the statement and the older group “strongly agrees”, but no differences are found between translations and post-editings within each group.

Apart from enjoyment, one direct question (“Was the excerpt interesting?”) with seven different options (from “very interesting” to “very boring”) aimed to assess their interest in the film. Overall results show that the translation was better evaluated than the post-editing (“translation median = “very interesting”, post-editing median= “pretty interesting”), although differences are found in the two excerpts under analysis: in excerpt 1 the translation is considered by all participants as either “very” or “pretty interesting”, whilst the post-editing is only considered as “very” or “pretty interesting” by 67.87% of participants. It is even qualified as “boring” by 10.71% of the participants. The difference is minimal though, as the median in both cases is "pretty interesting" for excerpt one. In the second excerpt, the trend changes: 82.14% consider the translation “very” or “pretty interesting”, whilst 100% qualify the post-editing as such. The difference in this case is higher, as the median is “very interesting” for post-editings and “pretty interesting” for translations. These are unexpected results since the dubbing studio

Table 4. Agreement level on interest (medians)
considered the second excerpt post-editing to be of low quality. When analysing the data according to the age groups, it can be observed that the 40 and over group prefer the translation (85.71% rated it as “pretty interesting” and 14.29% as “very interesting” while the younger group like the post-editing better (100% rated it as “very interesting”). To gather more information on interest, participants were also asked whether they would be willing to look for more information on the documentary, and the median reply in all conditions, regardless of excerpt, age and condition, was “maybe” (the middle option on a 7-point Likert scale). Similarly, to the question “Would you like to watch the whole documentary film?”, a positive reply was obtained in all conditions (median= “yes”), regardless of age. The only difference is found in the second excerpt, where those who watched the translated version react more positively (median= “yes”) than those that watch the post-edited (median = “maybe”).

Finally, when asked which of the two versions was their preferred one, without knowing which one was a post-editing or a translation, results show almost no difference between both versions: while 44.64% of the participants prefer a translated version, 42.86% prefer a post-edited one. However, when excerpts are analysed separately, it can be seen that participants prefer the translated version (50%) to the post-edited (35.71%) for the first excerpt, and the post-edited (50%) to the translated (39.26%) for the second. Differentiating between age groups, older viewers prefer the translated versions of both excerpts to the same extent (85.71%), whereas younger viewers prefer the post-edited version of the second excerpt (85.71%) and the translated version of the first (78.57%).

Overall results show slightly better results in some aspects for the translation (enjoyment, interest, and preferences), although different trends are observed when analysing the data independently for excerpts and age groups.

5.3.2. End-Users Comprehension

All participants considered the narration to be completely understandable and did not perceive any mistakes. However, results show slight differences in comprehension in some instances. Taking into account both excerpts and all participants, translated versions are better understood (mean score: 0.71) than post-edited ones (mean score: 0.66). When analysing each excerpt separately, opposite trends are observed: the translation is better understood in the first excerpt (translation= 0.79, post-editing= 0.63), whilst the post-editing is slightly better understood than translation in the second one (translation= 0.63, post-editing= 0.69). Considering both age groups, the younger group seems to understand better translated versions (translation= 0.72, post-editing= 0.61), whilst the older group obtains almost identical results (translation= 0.70, post-editing= 0.71).

In conclusion, results show slightly higher comprehension levels for the translation when considering all the data. Translation is also slightly higher in comprehension for the first excerpt and the younger group. Almost identical results are found for the older group, and slightly higher results in favor of post-editing are encountered for the second excerpt.
6. Conclusion and Further Research

This article presents an experiment in which the quality of post-edited wildlife documentary films is compared to the quality of translated documentaries in order to determine whether there is a meaningful difference between the qualities of each. Compared to other QA performed in the field of Translation Studies and PE, the QA used in this experiment was carried out in three levels: it takes into account not only experts but also end-users and the dubbing studio where the written script is converted into an oral recording.

The results of the study indicate that, according to experts, translations seem to perform better in the three evaluation rounds when global percentages are considered, but median results show no differences. A lower number of corrections is also performed on translations, although the differences are low. On the contrary, post-editings are better graded in more aspects than translations in the questionnaire after round 1, although differences are again minimal. And, finally, post-editings are more difficult to identify as such, which may be considered an indicator that no meaningful quality differences are observed.

When observational data from the dubbing session are analysed, translation also seems to perform better, although the differences in the first excerpt are minimal and higher in the second one.

Finally, when taking into account end-users, better median results are obtained for the translation in terms of enjoyment, interest, and user preferences, although a meticulous analysis of each excerpt and group yields diverging trends. It must be stressed, though, that the differences are low, and the same results are obtained for both conditions in the other items under analysis. In terms of comprehension, translation is better understood than post-editing when taking into account all the data, but also in the first excerpt and in the younger group. However, results are non-meaningful.

All in all, translation seems to receive better marks, although the difference is not high, and hence, not meaningful, proving our initial hypothesis.

When comparing the evaluation at the three stages, it can be inferred that expectations of end-users are not high, as their ratings were high compared to the rather low evaluations of both experts and the dubbing studio professionals. The low quality of both translations and post-editings might be due to the lack of experience of the MA students and the test conditions (volunteer work rather than professionally paid commission), which is a limitation of our research. It remains to be seen whether professional translators, with or without post-editing experience, would yield different results.

This study is limited in scope but it hopefully will open the door to future research in the field of audiovisual translation evaluation and post-editing. Future studies could take into account other language pairs, work with longer excerpts, and involve professional translators as well as experts in post-editing. Another stakeholder could be included in the evaluation, namely the broadcaster commissioning the VO of non-fictional genres. It may well be that quality expectations, and consequently evaluations of lecturers, professionals, broadcasters, dubbing directors, and end-users differ in many aspects, and analysing these different expectations is an interesting research topic. Additionally, a modified version of our experiment could include methodological improvements such as developing identical questions at different levels in order to obtain comparable data. We are fully aware that our research can be improved and expanded in many ways, but it has hopefully contributed to shed some light on an under-researched topic.
References


The Impact of Machine Translation Error Types on Post-Editing Effort Indicators

Joke Daems  
joke.daems@ugent.be  
Department Translation, Interpreting and Communication, Ghent University, Ghent, 9000, Belgium

Sonia Vandepitte  
sonia.vandepitte@ugent.be  
Department Translation, Interpreting and Communication, Ghent University, Ghent, 9000, Belgium

Robert Hartsuiker  
robert.hartsuiker@ugent.be  
Department of Experimental Psychology, Ghent University, Ghent, 9000, Belgium

Lieve Macken  
lieve.macken@ugent.be  
Department Translation, Interpreting and Communication, Ghent University, Ghent, 9000, Belgium

Abstract

In this paper, we report on a post-editing study for general text types from English into Dutch conducted with master’s students of translation. We used a fine-grained machine translation (MT) quality assessment method with error weights that correspond to severity levels and are related to cognitive load. Linear mixed effects models are applied to analyze the impact of MT quality on potential post-editing effort indicators. The impact of MT quality is evaluated on three different levels, each with an increasing granularity. We find that MT quality is a significant predictor of all different types of post-editing effort indicators and that different types of MT errors predict different post-editing effort indicators.

1. Introduction

In recent years, machine translation (MT) and its subsequent post-editing have become more widely accepted in the translation industry. Especially when it comes to technical texts, machine translation has proven its worth, with companies like Autodesk reporting on productivity increases in comparison with human translation ranging from 20 to 131%, depending on the language combination and translator (Plitt & Masselot, 2010). The main goal of post-editing research is no longer finding out whether or not post-editing can be used, but rather finding out when it cannot be used, and how machine translation systems can be improved to better suit post-editors’ needs.

While post-editing is generally assumed to be faster than human translation, speed is not the only factor that should be taken into account when assessing the post-editing process. More recent studies have looked at ways of determining post-editing effort. This knowledge can be used, on the one hand, to improve the quality of MT systems, and, on the other hand, to reduce post-editors’ frustration by only presenting them with a segment containing MT output when the effort required to post-edit that segment is not too high.

Krings (2001) mentioned three levels of post-editing effort: temporal effort, or the time needed to post-edit a given text, cognitive effort, or the activation of cognitive processes dur-
In post-editing, and technical effort, or the technical operations such as insertions and deletions that are performed during post-editing. According to Krings (2001), post-editing research should concentrate on causes and manifestations of post-editing effort with a focus on cognitive effort: "The type and extent of cognitive processes triggered by the post-editing task must be defined qualitatively and quantitatively, and correlated to the corresponding deficiencies in machine translations as triggering factors" (p. 182).

In this paper, we will first discuss some previous work on the effort involved in post-editing and the problems that arise when trying to measure cognitive effort in isolation. We then present the results of our study, examining the impact of different types of machine translation errors on post-editing effort indicators with student translators post-editing from English into Dutch.

2. Related research

The ultimate goal of post-editing process research is predicting how much effort a post-editor will need to correct a segment before presenting the post-editor with that segment. Depending on the expected effort, a translator can then be given MT output to post-edit whenever the effort to post-edit would be lower than the effort needed when translating that segment from scratch. Two aspects need to be researched in order to reach that ultimate goal: firstly, we need to establish which types of effort we take into account and how we can objectively measure them, and secondly, we need to find ways of predicting effort on the basis of elements contained in either the source text or the MT output. Both aspects will be discussed in the following paragraphs.

A number of potential post-editing effort indicators have been introduced in previous research. The distinction between temporal, cognitive and technical effort as proposed by Krings (2001), however, does not seem to be a clear distinction.

While temporal effort seems the easiest to measure, as it is simply the time needed to translate a word, segment or text, Koponen et al. (2012) found evidence that post-editing time can also be an indication of cognitive effort. They use a cognitively motivated MT error classification created by Temnikova (2010), but finish their paper with a few remarks on the classification and a suggestion for future work: "A revised set of error categories with more detailed error types (...) is also an interesting direction to help understand the cognitive load in post-editing" (p. 20). Koponen et al. (2012) also looked at a technical effort indicator - keystrokes - and its relationship to cognitive load. However, they found that keystrokes were influenced more by individual differences between participants than by cognitive load. We therefore decided not to include keystrokes as such in our analysis. Related to keystrokes are production units, or sequences of coherent typing activity. Although producing translation output in itself is clearly a technical activity, Lacruz et al. (2012) intuitively felt that an increase in the number of complete editing events (which correspond to the notion of production units) would lead to an increase in cognitive demand as well, making it a cognitive effort indicator in addition to a technical effort indicator. The question remains whether editing events really correspond to cognitive effort. For example, many spelling errors or adjective-noun agreement errors will require quite a few (local) editing events, but are not really difficult to solve.

Lacruz et al. (2012) further introduce the average pause ratio (the average time per pause in the segment divided by the average time per word in the segment) as an answer to O'Brien's pause ratio (2006) - the total time in pauses divided by the total editing time. O'Brien (2006) did not find conclusive evidence for a relationship between pauses and cognitive activity. Lacruz et al. (2012) argue that pause ratio is not sensitive enough as a measure for
cognitive activity, as it does not take average pause length into account. We include both pause measures in our study, to establish whether or not they can both be used, and whether or not they are indicators for different causes of effort. Lacruz et al. (2012) found a relationship between average pause ratio and the number of production units. As production units are delimited by pauses, and the average pause ratio is influenced by the number of pauses, perhaps this finding is related more to intrinsic correlation than to actual impact of cognitive load on pause behavior, although the relationship is most likely more complex. We will look at production units and average pause ratio in isolation to better understand the differences and similarities between both variables.

Some of the few effort indicators that seem to be exclusively related to cognitive post-editing effort, are the average fixation duration and the number of fixations. Building on the eye-mind hypothesis from Just and Carpenter (1980), a person is cognitively processing what they are looking at. Longer fixations should thus be an indication of more cognitive processing. This assumption was confirmed by Jakobsen and Jensen (2008), who found longer average fixation durations and a higher number of fixations as the complexity of the task increased from reading to translation. Doherty and O’Brien (2009), however, found a higher number of fixations for bad MT output than for good MT output, but they did not find a significant difference between the average fixation durations for both types. We will include both average fixation duration and number of fixations as potential cognitive post-editing effort indicators.

From the abovementioned research, it becomes clear that the distinction between the different types of effort indicators is not always easily made. Correlations are identified between different indicators without really knowing whether or not they measure different things. To avoid this circular thinking, we need to find a way of studying the post-editing effort indicators in isolation, by linking them to source text and MT output characteristics rather than to other post-editing effort indicators. O’Brien (2004) has taken a step in this direction by looking at negative translatability indicators (NTIs) in the source texts, or elements that can reasonably be considered to be problematic for MT systems, for example, long noun phrases or gerunds. Although some NTIs indeed seem to have an impact on post-editing effort, there are some NTIs that have no effect, and O’Brien (2004) further found post-editing activity in segments that did not contain NTIs. From these findings, we can derive that NTIs do not conclusively predict post-editing effort, and perhaps another focus is needed.

In this paper, we take a look at a fine-grained MT quality assessment and whether or not the average MT error weight of a segment has an impact on the post-editing process. In line with previous research, we take a look at different types of post-editing effort indicators.

3. Methodology

3.1. Participants

Participants were ten master's students of translation. All of them were native speakers of Dutch. They had no previous experience post-editing and had passed their final English General Translation exam. They received two gift vouchers of 50 euros each. As we are working with students, it is of course hard to say whether our results will generalize to the professional translation process. However, we have repeated the experiment with professional translators (but the process data has not yet been analyzed), and we found no significant differences in proficiency or attitude towards post-editing between the two groups, so perhaps they are more comparable than often thought.
3.2. Text selection

The present study is a part of a larger study aimed at comparing the differences between the human translation process and the post-editing process for students and professional translators for general text types. In the present study, the focus will be on the post-editing process of the students only, but the texts have been selected with the larger study in mind.

Originally, fifteen different English newspaper articles were selected from newsela.com, a website providing newspaper articles at different levels of complexity, as indicated by a Lexile score. We selected articles with the same level of complexity, i.e., Lexile scores between 1160L and 1190L\(^1\), to try to control for textual differences in our studies. Each article was reduced to its first 150-160 words, and then analyzed for additional readability measures and potential translation problems. Texts with on average less than fifteen or more than twenty words per sentence were discarded, as well as texts that contained too many or too few complex compounds, idiomatic expressions, infrequent words or polysemous words. Sentence length ranged from seven to thirty-five words, with an average of eighteen point three, and a median of eighteen words per sentence. The texts were then translated into Dutch by the statistical machine translation system Google Translate. We annotated the MT output for quality, as will be discussed in section 3.4. From the original fifteen texts, the eight texts that were most comparable in difficulty - based on the potential translation problems and MT output quality - were retained. Texts have different subjects and don’t require specialist knowledge to be translated.

3.3. Experimental setup

Two sessions were planned for each participant. During the first session, students had to first fill out a survey, and take a LexTALE test (Lemhöfer & Broersma, 2012) to be able to take their English proficiency into account. This was followed by a copytask and a warmup task combining both post-editing and human translation, so the students could get used to the tools and different types of tasks. After the warmup, students post-edited two texts and translated two texts from scratch. During the second session, students again started with a warmup task, followed by two human translation tasks and two post-editing tasks. The order of texts and tasks was balanced in a Latin square design across all participants, to reduce task order effects. The second session ended with a retrospective part, during which students could highlight elements in the text that they found most difficult to translate or post-edit, and another survey to measure how students experienced the experiment and the different tasks.

To be able to look at different aspects of post-editing effort, we used a combination of keystroke logging tools and eye tracking. The process itself was registered by the CASMACAT translator’s workbench (Alabau et al., 2013), which looks like an actual translation environment to improve ecological validity, yet contains keystroke logging and mouse tracking software for researchers to better be able to observe the translation and post-editing process in detail. The texts were presented to the students one by one, and each text was subdivided in translation segments, corresponding to sentences in the source text. The number of segments in each text ranges from seven to ten. A plugin connects Casmacat to the EyeLink 1000 eyetracker that was used to register the students’ eye-movements while translating and post-editing. In addition to these tools, an extra keystroke logging tool, Inputlog (Leijten & Van Waes, 2013) was running in the background. While the CASMACAT software is capable of performing a detailed logging within the CASMACAT interface, it cannot log external applications. Inputlog registers when and which applications other than CASMACAT are

\(^1\) The authors would like to thank MetaMetrics® for their permission to publish Lexile scores in the present paper. https://www.metametricsinc.com/lexile-framework-reading
opened, and which keystrokes are performed within those screens. Though not applicable for
the present study, this information can lead to better insights regarding translators’ usage of
external resources.

In total, we collected forty student post-editing sessions and forty student human trans-
lation sessions. In this paper, we’ll focus on the post-editing sessions only. Each of the eight
texts was post-edited by five different students. For some segments, some of the data was
missing, so these segments were left out of the analysis. The final dataset consisted of 317
post-edited segments.

3.4. MT quality annotation

MT quality can be measured in a myriad of ways, depending on the goal of the assessment
and the means available. Automatic metrics like BLEU (Papineni, Roukos, Ward, & Zhu,
2002) are often used to evaluate the output of MT systems by comparing it to reference trans-
lations. While these metrics give an indication of an MT system’s performance, they rely on
the idea that MT quality in itself should approach human quality. Metrics like human-targeted
translation error rate (HTER)(Snover et al., 2006), focus more on the perspective of post-
editing: how much editing effort is needed to make the MT output match a reference transla-
tion? This is the difference between judging the quality of MT output as a final product, and
judging the utility of MT output for subsequent post-editing, which is discussed in more detail
by Denkowski and Lavie (2012). While both types of metric have proven their worth for dif-
ferent applications, they depend on the availability of human reference translations, which is
not something that is always readily available. Ideally, MT evaluation can take place without
resorting to reference translations, so it can be used on new texts as well. The translation qual-
ity assessment approach presented in this paper builds on that notion, while at the same time
being flexible enough so it can be used to evaluate human translation quality and post-editing
quality as well (while not relevant for this particular study, it is of importance to the larger
study we are conducting).

The quality of MT output is judged from two different perspectives. On the one hand, there is
the adherence to the target text and target language norms, also known as acceptability, and,
on the other, the adherence to the source text norms and meaning, also known as adequacy
(Toury, 1995). This distinction has been used in context of human translation evaluation by
Koby and Champe (2013), with acceptability and adequacy issues being called mechanical
and transfer errors, respectively. In a more recent study by Koponen and Salmi (2015), where
participants had to correct MT output without the source text, the distinction was used suc-
cessfully as well. Yet the researchers felt the need for a more fine-grained error analysis to
better establish which MT errors are the most difficult to edit, and which MT errors lead to
meaning loss in the final post-edited text. Koponen and Salmi use Temnikova’s (2010) MT
error classification and cognitive ranking, but identify a few shortcomings of the ranking,
especially with regards to punctuation errors. Lacruz et al. (2014) propose another MT error
classification. They use the ATA grading rubric (Koby & Champe, 2013) to distinguish be-
tween mechanical and transfer errors, and also create their own classification, which is a sim-
plified version of the ATA’s rubric. Lacruz et al. (2014) expect that “cognitive demand placed
on post-editors by transfer errors is greater than the cognitive demand resulting from mech-
anical errors” (p. 77). Mechanical and transfer errors correspond roughly to acceptability and
adequacy errors as discussed below, although this is somewhat of an oversimplification. Fol-
lowing Lacruz et al.’s definition (2014), mechanical errors are those errors that can be solved
without looking at the source text, whereas our acceptability errors can be identified as errors
without looking at the source text, but they cannot necessarily be solved without consulting
the source text.
To be able to distinguish more clearly between the effects of acceptability and adequacy issues in MT, we suggest adopting a two-step translation quality assessment approach. In a first step, annotators only receive the target text (in this case, MT output) and they annotate all acceptability issues (grammar and syntax, coherence, lexicon, style and register, spelling). In a second step, annotators receive both the source and the target text, and they annotate all discrepancies in meaning between source and target text, i.e. adequacy issues. It is possible for issues to have an impact on adequacy as well as acceptability, in which case both errors will be annotated. For example, a word sense error is an adequacy issue that can lead to a logical problem in the target text, which in turn is an acceptability issue.

We have tested and fine-tuned our two-step translation quality assessment approach in two pilot studies with student translators, on two different text types (newspaper articles and technical texts), and have successfully applied the approach on MT output as well. To ensure as much objectivity and quality of the annotations as possible, two different people - both authors of this paper with a background in translation studies and evaluation - annotated all the texts. After the annotation process, the annotators discussed discrepancies in their annotations, and only the annotations that both annotators agreed on were kept for the final analysis. The annotations were made with the brat rapid annotation tool (Stenetorp et al., 2012).

To allow for a deeper analysis than hitherto possible, we created a very fine-grained analysis for acceptability and adequacy issues (for an overview of subcategories, see Daems, Macken, & Vandepitte, 2013). Though originally intended for translation evaluation of English to Dutch texts, the categorization builds on common evaluation metric categories and can easily be expanded to other languages. For example, when working with grammatically more complex languages, subcategories for cases could be added to the acceptability category 'grammar and syntax'.

In line with Temnikova (2010) and Lacruz et al. (2014), we believe that error categorizations need to incorporate some method of ranking the different errors according to severity. The errors in our categorization receive error weights ranging from 0 (no actual error, but can be interesting to annotate, such as explicitations), to 4 (critical problems, such as contradictions, that have a critical impact on the understandability of the text). Depending on the text type or task, the weights can be set differently. For example, in technical texts, terminology errors are critical errors, whereas they are not as dramatic in general texts. While we did not originally assign error weights with cognitivity in mind, but rather with the translation product in mind, we do see a close correspondence between the two aspects. For example, structural issues (error weight = 3) will be cognitively demanding to solve, but they also make the text as a product harder to understand. Likewise, capitalization errors (error weight = 1) are easy to solve, and they hardly impair the understanding of the text.

3.5. MT data

Of the 63 source text segments, only three segments contained no errors in the MT output. In total, 201 acceptability issues were identified, and 86 adequacy issues. Though the original error categorization is really fine-grained to allow for detailed analysis, the current dataset is a bit too small to be able to perform any statistical tests on such a detailed level. We therefore decided to group some of the categories together into higher-order categories, so that each category appeared at least ten times in the dataset. The final classification can be seen in Figure 1.

All subcategories for style and spelling have been grouped together into the main categories, since there were very few instances of these subcategories. For lexicon, the subcategory 'wrong collocation' appeared often enough to stand alone, the other subcategories (wrong preposition, named entity, and word non-existent) have been grouped into 'lexicon other'. For
coherence issues, the category 'logical problem' occurred more than ten times, but the other categories together (conjunction, missing info, paragraph, and inconsistency) did not occur more than ten times, so all coherence categories were grouped together. The category 'adequacy' in Figure 1 contains all forms of adequacy issues. Mistranslations and word sense issues occurred frequently enough to be considered as separate categories, the other subcategories (additions, deletions, misplaced words, function words, part of speech, inconsistent terminology) were grouped together into 'adequacy other'. Within the grammar and syntax category (the most common error category for MT output), word order issues, structural issues and incorrect verb forms occurred more than ten times each. The different types of agreement issues (noun-adjective, article-noun, subject-verb, and reference) were grouped into a new 'agreement' category, and the other grammatical issues are contained in the 'grammar other' category (superfluous or missing elements).

Figure 1. Overview of regrouping and number of occurrences of each error type in the MT output.

4. Analysis

The main goal is to analyze the effect of machine translation quality on different types of post-editing effort indicators. We looked at MT quality on three different levels. For the first level, we simply look at the effect of the average total error weight per word on the different effort indicators. As discussed above, some errors lead to adequacy problems as well as acceptability problems. In these cases, only the adequacy error weight was taken into account for the calculation of the total error weight, as the acceptability error was caused by the adequacy error. For the second level, we look at the impact of the average acceptability and adequacy error weight per word on the different post-editing effort indicators. For the third level, we go even more fine-grained, and we identify the different subcategories that are best suited to predict changes in the post-editing effort indicators.

Based on previous research, we look at the following post-editing effort indicators as dependent variables:

- **Average number of production units**: technical and/or cognitive effort (see discussion of Lacruz et al., 2012 in Section 2), calculated by dividing the number of production units of a segment by the number of source text words in that segment
• **Average time per word**: temporal and/or cognitive effort (see discussion of Koponen et al., 2012 in Section 2), calculated by dividing the total editing time (in ms) of a segment by the number of source text words in that segment

• **Average fixation duration**: cognitive effort (Just and Carpenter, 1980), calculated by dividing the total fixation duration (in ms) of a segment by the number of fixations within that segment

• **Average number of fixations**: cognitive effort (Doherty and O'Brien, 2009), calculated by dividing the number of fixations in a segment by the number of source text words in that segment

• **Pause Ratio**: technical and/or cognitive effort, as suggested by O'Brien (2006), calculated by dividing the total time in pauses (in ms) for a segment by the total editing time (in ms) for that segment

• **Average Pause Ratio**: technical and/or cognitive effort, as suggested by Lacruz et al. (2012), calculated by dividing the average time per pause in a segment by the average time per word in a segment

The dependent variables were derived from the SG-data files obtained by processing the output from CASMACAT. Production units are sequences of coherent typing activity, separated from one another by pauses of at least 1000 ms. A segment in CASMACAT corresponds to a sentence in the source text. A pause in CASMACAT corresponds to any pause in typing activity lasting at least 1000 ms - the lowest possible pause threshold present in the CASMACAT output. The fixation information is added to the CASMACAT tables via the EyeLink plugin.

4.1. **Level 1: Average total MT error weight**

We used the R statistical software (R Core Team, 2014) to perform linear mixed effects analyses with the lme4 package (Bates et al., 2014) and the lmerTest package (Kuznetsova et al., 2014). For each of our independent variables, we built a null model without fixed effect, but with sentence code and participant as random factors, to account for between text and between participant variation. The only exception was the dependent variable 'average fixation duration', for which the output from the step-function from the lmerTest package showed that only participant was necessary as a random factor. We then tested this null model against a model with fixed effect: the average total MT error weight. As can be seen in Table 1, the model with fixed effect was always significantly different from the null model without fixed effect, with p ranging from < 0.001 (when predicting the average number of production units or the average pause ratio) to 0.017 (when predicting the average fixation duration). Likewise, the Akaike's Information Criterion (AIC) value is always lower for the model with predictor. AIC (Akaike, 1974) is a method designed for model selection, based on a comparison between models. According to Burnham and Anderson (2004), the best model is the model with the lowest AIC value. Their rule of thumb states that if the difference between models is less than 2, there is still substantial support for the weakest model. If the difference is between 4

---

2 We would like to thank one of the anonymous reviewers for pointing out that the measure of average pause ratio in this paper is somewhat different from that presented in Lacruz et al. (2012), since the pause threshold is set at 1000ms, whereas Lacruz et al. included clusters of shorter pauses as well. This needs to be taken into account when studying these findings.
and 7, there is far less support for the weakest model, and if the difference is greater than 10, there is hardly any support for the weakest model. As can be seen in Table 1, the difference in AIC values ranges from 17 to 4, always in favor of the model with average total MT error weight as predictor variable. It must be noted that the AIC value in itself has no meaning. The values are used for comparison between models predicting the same dependent variable, but cannot be compared across models predicting different dependent variables.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Random factors</th>
<th>AIC without predictor</th>
<th>AIC with predictor</th>
<th>effect</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of production units</td>
<td>sentence code, participant</td>
<td>-78</td>
<td>-95</td>
<td>0,3 (± 0,06)</td>
<td>&lt; 0,001</td>
</tr>
<tr>
<td>Average duration per word (in ms)</td>
<td>sentence code, participant</td>
<td>5979</td>
<td>5974</td>
<td>3077 (± 1153)</td>
<td>0,01</td>
</tr>
<tr>
<td>Average fixation duration (in ms)</td>
<td>participant</td>
<td>2890</td>
<td>2886</td>
<td>12 (± 5)</td>
<td>0,017</td>
</tr>
<tr>
<td>Average number of fixations</td>
<td>sentence code, participant</td>
<td>2268</td>
<td>2262</td>
<td>8,6 (± 3)</td>
<td>0,005</td>
</tr>
<tr>
<td>Pause Ratio</td>
<td>sentence code, participant</td>
<td>-688</td>
<td>-698</td>
<td>-0,07 (± 0,02)</td>
<td>0,002</td>
</tr>
<tr>
<td>Average Pause Ratio</td>
<td>sentence code, participant</td>
<td>1596</td>
<td>1580</td>
<td>-3,86 (± 0,85)</td>
<td>&lt; 0,001</td>
</tr>
</tbody>
</table>

Table 1. Summary of mixed models with average total MT error weight as fixed effect.

The impact of the predictor variable can be derived from the 'effect' column in Table 1. The column should be read as follows: for each increase of the average MT error weight per word by one unit, the corresponding dependent variable changes by the value in the 'effect' column. For example, for each unit increase in the average MT error weight per word, the average duration per word increases with 3 seconds. All of the models show the expected effects. A decrease in MT quality, i.e., an increase in average MT error weight, leads to an increase of the number of production units, the average duration per word, the average fixation duration, and the average number of fixations, and to a decrease of the pause ratio and average pause ratio. The latter is in line with findings by Lacruz et al. (2012) that high cognitive load is related to lower average pause ratios. It's remarkable as well that we find a small but statistically significant effect of MT quality on pause ratio. O'Brien (2006) looked at the effect of negative translatability indicators on pause ratio and did not find a statistically significant difference between sentences with NTIs and with few or no NTIs. We can assume that MT error weights provide a more accurate estimation of pause behavior than NTIs, although the direction of the effect is somewhat surprising. Following O'Brien (2006), higher cognitive load should be associated with a higher pause ratio, which is in contrast with the effect we see in Table 1. More detailed analysis is needed to further examine this effect. Our findings for the average fixation duration seem to support the findings by Jakobsen and Jensen (2008) that increased task complexity leads to longer fixations. It must be noted, however, that - though statistically significant - the observed change of 12 ms in our study is rather small and perhaps...
not practically significant. The change in average number of fixations seems to be a more convincing effort indicator, as was found by Doherty and O’Brien (2009).

4.2. Level 2: Average acceptability and adequacy MT error weight

In the second step, we used the same null models as in the step above, but we applied a more fine-grained approach. Both average acceptability and adequacy error weight were added as possible predictor variables to the model. The step-function from the lmerTest package was used to find out which elements needed to be retained in the final model. This step-function showed that adding adequacy as a predictor did not significantly improve the model, so the final model consisted of only average acceptability error weight as a predictor. The random effects were the same as in the previous analysis. A summary of the models can be found in Table 2.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Predictor variable(s)</th>
<th>AIC without predictor</th>
<th>AIC with predictor</th>
<th>effect</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of production units</td>
<td>acceptability</td>
<td>-78</td>
<td>-92</td>
<td>0.32 (± 0.07)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Average duration per word (in ms)</td>
<td>acceptability</td>
<td>5979</td>
<td>5973</td>
<td>3347 (± 1312)</td>
<td>0.013</td>
</tr>
<tr>
<td>Average fixation duration (in ms)</td>
<td>acceptability</td>
<td>2890</td>
<td>2884</td>
<td>15 (± 5)</td>
<td>0.007</td>
</tr>
<tr>
<td>Average number of fixations</td>
<td>acceptability</td>
<td>2268</td>
<td>2264</td>
<td>8.5 (± 3.4)</td>
<td>0.015</td>
</tr>
<tr>
<td>Pause Ratio</td>
<td>acceptability</td>
<td>-688</td>
<td>-697</td>
<td>-0.08 (± 0.02)</td>
<td>&lt; 0.002</td>
</tr>
<tr>
<td>Average Pause Ratio</td>
<td>acceptability</td>
<td>1596</td>
<td>1580</td>
<td>-4.35 (± 0.97)</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Table 2. Summary of mixed models with average total adequacy and acceptability error weight as potential fixed effects.

The fact that the average adequacy error weight is not retained in the model is a bit counterintuitive, as Lacruz et al. (2014), for example, found that transfer errors (which roughly correspond to adequacy errors) were cognitively more demanding than mechanical errors. A possible explanation can be that transfer errors and mechanical errors are not entirely comparable to adequacy and acceptability errors, respectively. From Table 2, we can derive that average MT acceptability error weight is a significant predictor for all the different post-editing effort indicators, with p-values below the 0.01 level, with the exception of the p-values for dependent variables 'average duration per word' and 'average number of fixations', although the values are still well below the generally accepted 0.05 significance threshold.

The AIC values are somewhat different from the AIC values of the models with average total MT error weight as predictor variable, but the difference is never greater than three, so we can assume that both models can be supported. Again, we observe the same trend as with the average error weight per word. An increase in average acceptability error weight, leads to an increase of the number of production units, the average duration per word, the
average fixation duration, and the average number of fixations, and to a decrease of the pause ratio and average pause ratio.

4.3. Level 3: Average MT error weight for all subcategories

In a final step, we wanted to get a better idea of exactly which types of machine translation errors best predict the different types of post-editing effort indicators. We again used the post-editing effort indicators as dependent variables, and sentence and participant as random factors. This time, however, we added the average MT error weight for all the different subcategories to the model as potential predictor variables: mistranslation, word sense, adequacy other, coherence, style, lexicon other, wrong collocation, spelling, grammar other, structure, verb form, word order, and agreement (see Figure 1). The step-function was used to identify the variables that significantly added to the model. Only these variables were added to the final model, of which the results can be seen in Table 3.

The column 'predictor variables' gives an overview of the different subcategories that predict a change in the dependent variable. Comparing the AIC values of the model with predictors as shown in Table 3 with those from Table 2, we can see that there is more support for the fine-grained model than for the model with the average total MT error weight as a predictor for average duration per word, average fixation duration, and pause ratio. The opposite is true for average number of production units, and average pause ratio. It must be noted that AIC penalizes models with more predictor variables, and seeing how both the model predicting production units and the model predicting average pause ratio contain three or four predictor variables (in comparison with only one or two for the other post-editing effort indicators), this is not such a surprising fact.

What is interesting, is how the different post-editing effort indicators are influenced by different MT error types. The pause ratio and average pause ratio seem to be predicted by a subset of the variables that are predictors for the average number of production units. This is in line with our hypothesis that production units and pauses are closely related to one-another (seeing how the boundaries of production units are defined by pauses). Our findings are comparable to those of Lacruz et al. (2014), who found a strong correlation between pause to word ratio (an alternative for average pause ratio) and mistranslations and structural issues. They also found a correlation with insertions and deletions that we did not find in our data. This can be explained by the fact that insertions and deletions rarely occurred in our data (three and six times respectively), and perhaps their effect is nullified by grouping them together with other categories. The surprisingly negative effect of error weight on pause ratio as found in Table 1, might be explained by the types of errors found in Table 3: grammatical errors and word order errors are easily spotted, and also easily corrected. This would imply that sentences containing a lot of grammatical or word order errors require fewer time in pauses than sentences containing other types of errors, since these errors can be solved immediately. The average duration per word is predicted most by average MT error weight for coherence and structure issues, which indeed take a lot of time to solve: coherence issues require a post-editor to figure out how the text is semantically structured, whereas structural issues often contain a combination of grammatical structures, so that there are different ways of solving the problem, leading to a higher cognitive load and thus processing time. Fixation duration can be predicted by the average MT error weight for mistranslations, which can be explained by the fact that a mistranslation draws the attention and the problem is often harder to understand than is the case with, for example, grammatical errors. The average number of fixations can be predicted by the average MT error weight for coherence issues. Solving coherence issues requires a post-editor to look back and forth throughout the text to figure out how everything is connected, and so more fixations are needed.
Table 3. Summary of mixed models with average MT error weight for the subcategories retained by step function as fixed effects and sentence code and participants as random factors.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Predictor variables</th>
<th>AIC without predictors</th>
<th>AIC with predictors</th>
<th>effect</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of production units</td>
<td>mistranslation</td>
<td>-78</td>
<td>-91</td>
<td>0.32 (± 0.13)</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>grammar</td>
<td></td>
<td></td>
<td>0.34 (± 0.41)</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>structure</td>
<td></td>
<td></td>
<td>0.58 (± 0.22)</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>word order</td>
<td></td>
<td></td>
<td>0.43 (± 0.19)</td>
<td>0.028</td>
</tr>
<tr>
<td>Average duration per word (in ms)</td>
<td>coherence</td>
<td>5979</td>
<td>5964</td>
<td>6365 (± 2464)</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>structure</td>
<td></td>
<td></td>
<td>8020 (± 3912)</td>
<td>0.044</td>
</tr>
<tr>
<td>Average fixation duration (in ms)</td>
<td>mistranslation</td>
<td>2890</td>
<td>2882</td>
<td>30 (± 9)</td>
<td>0.002</td>
</tr>
<tr>
<td>Average number of fixations</td>
<td>coherence</td>
<td>2268</td>
<td>2264</td>
<td>16.4 (± 6.5)</td>
<td>0.015</td>
</tr>
<tr>
<td>Pause Ratio</td>
<td>grammar</td>
<td>-688</td>
<td>-700</td>
<td>-0.15 (± 0.04)</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>word order</td>
<td></td>
<td></td>
<td>-0.13 (± 0.06)</td>
<td>0.036</td>
</tr>
<tr>
<td>Average Pause Ratio</td>
<td>mistranslation</td>
<td>1595</td>
<td>1587</td>
<td>-4.52 (± 1.74)</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>structure</td>
<td></td>
<td></td>
<td>-6.27 (± 3.04)</td>
<td>0.043</td>
</tr>
<tr>
<td></td>
<td>word order</td>
<td></td>
<td></td>
<td>-6.1 (± 2.66)</td>
<td>0.025</td>
</tr>
</tbody>
</table>

5. Conclusion and discussion

To be able to predict post-editing effort, we need to look at source text and MT output features as possible influencing factors of the post-editing process. In this paper, we investigated how translation students' post-editing process was influenced by the average error weight of the MT output. We found that average MT error weight is a good predictor of six different post-editing effort indicators (average number of production units, average duration per word, average fixation duration, average number of fixations, average pause ratio, and pause ratio). The analysis was conducted on three levels of granularity, by means of linear mixed effects models. With regards to the more fine-grained level, we found that the different post-editing effort indicators are predicted by different MT error categories, with mistranslations, structural issues and word order issues being the most common categories. The average number of production units and the (average) pause ratio seem to be linked, as they are best predicted by comparable MT error categories, consisting of more technical fixes. Cognitively more demanding fixes (coherence and mistranslation) are better predictors for other types of post-editing effort indicators (average fixation duration, average number of fixations, and average duration per word).

We only looked at a few potential post-editing effort indicators and only at MT quality as a possible cause, but there are of course many more indicators and potential causes that can
help us better understand the post-editing process. In the future, we would like to look at syntactical variety between source and target language and translation entropy (Carl & Schaeffer, 2014). Other directions for future research include a more fine-grained analysis. We now looked at the segment level, but it could be interesting to look at production units and pauses in isolation, as there is usually a very long pause at the beginning of a segment before the first edit that might influence the pause data. In addition, we want to compare fixations on the source and target text, and focus more on specific MT errors rather than on the entire dataset at once.

Though we only had time to analyze the students' data, we conducted the same experiment with professional translators, and it will be interesting to compare our current findings with the results from the professional translators' data.
References


Living on the edge: productivity gain thresholds in machine translation evaluation metrics

Carla Parra Escartín  
Manuel Arcedillo  
Hermes Traducciones  
C/ Cóloquide, 6, portal 2, 3.º I  
28230 Las Rozas, Madrid, Spain

carla.parra@hermestrans.com  
manuel.arcedillo@hermestrans.com

Abstract  
This paper studies the minimum score at which machine translation (MT) evaluation metrics report productivity gains in a machine translation post-editing (MTPE) task. We ran an experiment involving 10 professional in-house translators from our company in which they were asked to carry out a real translation task involving MTPE, translation from scratch and fuzzy-match editing. We then analyzed the results and evaluated the MT output with traditional MT evaluation metrics such as BLEU and TER, as well as the standard used in the translation industry to analyze text similarity in translation memory (TM) matches: the fuzzy score. We report where the threshold between productivity gain and productivity loss lies and contrast it with past experiences in our company. We also compare the productivity of similar segments from MTPE and TM match editing samples in order to gain further insights on their cognitive effort and pricing schemes.

1 Introduction  
Over the past few years, translators are experiencing the introduction of MT in their translation workflow. However, it is often difficult for the parties involved to assess if the MT output quality allows any productivity gain and, if applicable, justifies any rate discount. A popular method for evaluating MT output involves using automatic metrics such as BLEU (Papineni et al., 2001) and TER (Snover et al., 2006). However, their estimation may prove technically difficult for general users and their results may be obscure to interpret in terms of productivity due to lack of conclusive research or lack of familiarity with these metrics.

To overcome these challenges, alternative metrics have been proposed which aim at applying the familiarity of translation memory (TM) fuzzy match scores (FMS) to MTPE evaluation as a target-side FMS (Parra Escartín and Arcedillo, 2015b). In this paper, we analyze BLEU, TER and FMS values obtained in an experiment involving ten professional translators with the aim of finding out if a threshold for productivity gain can be found in these metrics. We also discuss how cognitive effort and confidence in MT or TM technologies may impact post-editors’ productivity by comparing MTPE and TM match editing throughputs.

The remainder of this paper is structured as follows: Section 2 describes the experiment we carried out. Section 3 registers the productivity (cf. Subsection 3.1) and MT automatic evaluation metrics (cf. Subsection 3.2) obtained. In Section 4 we look at the threshold between productivity gain and productivity loss by correlating the results from Section 3. We then discuss the relation between productivity gains and cognitive effort in Section 5 before summarizing our research and discussing future work in Section 6.
2 Experiment settings

Ten in-house translators were asked to translate the same file from English into Spanish using memoQ, one of their most common computer-assisted translation (CAT) tools. This tool was chosen because it allows to keep track of the time spent in each segment. Translators were only allowed to use the TM, terminology database and MT output included in the hand-off package. We disabled all other memoQ’s productivity enhancing features, such as predictive text, sub-segment leverage and automatic fixing of fuzzy matches.

2.1 Test set

The text to be translated had over 7,000 words and was part of a software user guide. It belongs to a real translation request, except we filtered out repetitions and internal leverage segments to avoid skewing due to the inferior effort required to translate the second of two similar segments. All traditional TM fuzzy match bands, exact TM matches and no-match segments were represented in the text. Table 1 shows the word count distribution reported by memoQ.

<table>
<thead>
<tr>
<th>TM match</th>
<th>Words</th>
<th>Segments</th>
<th>Words/segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>1226</td>
<td>94</td>
<td>13.04</td>
</tr>
<tr>
<td>95-99%</td>
<td>231</td>
<td>21</td>
<td>11.00</td>
</tr>
<tr>
<td>85-94%</td>
<td>1062</td>
<td>48</td>
<td>22.12</td>
</tr>
<tr>
<td>75-84%</td>
<td>696</td>
<td>42</td>
<td>16.57</td>
</tr>
<tr>
<td>No Match</td>
<td>3804</td>
<td>263</td>
<td>14.46</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>7019</strong></td>
<td><strong>468</strong></td>
<td><strong>14.99</strong></td>
</tr>
</tbody>
</table>

Table 1: Word count as computed by memoQ.

We randomly divided the no-match band in two halves using the test set generator included in the m4loc package. One half was translated from scratch, without TM or MT suggestion, while the second half was sent to one of our custom MT engines.

2.2 Machine Translation system used

We used Systran to generate the raw MT output. This is the customized rule-based MT engine we normally use with this client. It can be considered a mature engine, since at the time of the experiment it had been subject to ongoing in-house customization for over three years and boasted a consistent record for productivity enhancement. Its customization includes dictionary entries, software settings, and pre- and post-editing scripts. Although Systran includes a statistical component for automatic post-editing, this was not used in our experiment.

2.3 Translators participating in the experiment

Ten professional translators were engaged in the experiment. Two carried out the task as part of a pilot experiment (Parra Escartín and Arcedillo, 2015a) which studied the feasibility of using a target-side FMS as an alternative MT evaluation metric. Eight additional translators were subsequently engaged to perform the same task under the same conditions with the aim of verifying the initial findings and provide deeper insights (Parra Escartín and Arcedillo, 2015b).

The translators were asked to provide their years of experience in translation and MTPE, their experience working in texts from this client, their opinion on MT (positive or negative),

---

1The version used was memoQ 2015 build 3.
2For further details on the experiment settings described in this section, see Parra Escartín and Arcedillo (2015b).
3https://code.google.com/p/m4loc
4Systran 7 Premium Translator was used.
and their opinion on CAT tools (positive or negative). Table 2 summarizes the results of our survey. Translators are sorted in descending order according to their combined experience in translation and MTPE. Two translators expressed that they did not like working with MT, despite acknowledging that high quality MT output generally enhances their productivity. This negative bias, however, did not seem to affect the results.

<table>
<thead>
<tr>
<th>Trans. exp.</th>
<th>MTPE exp.</th>
<th>Client exp.</th>
<th>MT opinion</th>
<th>CAT opinion</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR-1</td>
<td>5</td>
<td>3</td>
<td>Yes</td>
<td>Positive</td>
</tr>
<tr>
<td>TR-2</td>
<td>5</td>
<td>3</td>
<td>Yes</td>
<td>Positive</td>
</tr>
<tr>
<td>TR-3</td>
<td>5.5</td>
<td>2</td>
<td>Yes</td>
<td>Positive</td>
</tr>
<tr>
<td>TR-4</td>
<td>5</td>
<td>1</td>
<td>Yes</td>
<td>Negative</td>
</tr>
<tr>
<td>TR-5</td>
<td>5</td>
<td>0</td>
<td>No</td>
<td>Positive</td>
</tr>
<tr>
<td>TR-6</td>
<td>5</td>
<td>0</td>
<td>No</td>
<td>Positive</td>
</tr>
<tr>
<td>TR-7</td>
<td>2</td>
<td>1</td>
<td>Yes</td>
<td>Positive</td>
</tr>
<tr>
<td>TR-8</td>
<td>1.5</td>
<td>1.5</td>
<td>Yes</td>
<td>Positive</td>
</tr>
<tr>
<td>TR-9</td>
<td>2</td>
<td>0</td>
<td>No</td>
<td>Negative</td>
</tr>
<tr>
<td>TR-10</td>
<td>1</td>
<td>0</td>
<td>No</td>
<td>Positive</td>
</tr>
</tbody>
</table>

Table 2: Overview of translator’s experience (measured in years) and opinion on MT and CAT.

The translators were provided with a translation package to be opened in memoQ, where all TM and MT suggestions appeared as pre-translated text. They did not have to choose the type of output to post-edit in each segment: they were provided with either a TM or MT suggestion (or no suggestion at all in the case of the translation-from-scratch subsample). These segments were marked according to their standard treatment in memoQ and similar CAT tools: blue and their TM match score in the case of TM suggestions, and a different color and no TM match score in the case of MT suggestions. Whenever editing a TM match, the application would show the difference between the source text stored in the TM and the new source text.

3 Data analysis

The ten packages delivered by the translators were analyzed individually in order to extract the time spent in each segment and calculate automatic evaluation metrics. Even though translators were instructed to perform all necessary edits to achieve the usual quality required by our client, Translator 5’s sample showed clear signs of under-editing which would not have reached that goal. Meanwhile, it turned out that Translator 8 enabled memoQ’s predictive text feature, thus avoiding an adequate comparison with the rest of samples. This led us to discard the whole sets of Translators 5 and 8.

Of the remaining 3744 segments, we also discarded 6 due to too low or too high productivity (over 40 seconds per word and under 0.2 seconds per word, respectively). These outlier limits were found by computing interquartile ranges. The segments with too long editing times may be due to the translator not closing the editor during pauses or long breaks (in fact, the duration and time of at least two of those segments clearly match translator’s lunch breaks). As to the segments left out due to unnaturally high productivity, most of them (28 out of 29) correspond to 100% matches. An explanation for them might be that the translator actually revised the TM suggestion while having the cursor placed in a contiguous segment, and only entered the segment to confirm it. It may also be that translators did not even read those 100% matches.

5For further details on these issues, see Parra Escartín and Arcedillo (2015b).
6Federico et al. (2012) also establish outlier thresholds in their experiment. In their case, the upper threshold was 30 seconds per word, while the lower was 0.5 seconds per word.
and instead directly confirmed them, either because they trusted the TM suggestion or because they intended to revise them at a later stage during self-revision. However, we lack a way to verify these hypotheses.

3.1 Productivity report

Table 3 reports the results obtained for each translator in each band. The average words per hour across translators is provided in the last column, while the last row shows the productivity gain achieved by the MTPE band over translation from scratch. This MTPE gain is calculated according to Equation 1, where PE_Throughput is the productivity achieved by one translator in words per hour when post-editing MT output, and TRA_Throughput is the productivity achieved by that same translator when translating from scratch.

\[
Gain = \left( \frac{PE_{Throughput}}{TRA_{Throughput}} - 1 \right) \times 100
\]

(1)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>94</td>
<td>1226</td>
<td>3277</td>
<td>2942</td>
<td>1894</td>
<td>1767</td>
<td>1579</td>
<td>4039</td>
<td>2798</td>
</tr>
<tr>
<td>95-99%</td>
<td>21</td>
<td>231</td>
<td>2642</td>
<td>2625</td>
<td>1476</td>
<td>1299</td>
<td>963</td>
<td>2011</td>
<td>1133</td>
</tr>
<tr>
<td>85-94%</td>
<td>48</td>
<td>1062</td>
<td>2960</td>
<td>2248</td>
<td>1660</td>
<td>1678</td>
<td>1232</td>
<td>2164</td>
<td>2429</td>
</tr>
<tr>
<td>75-84%</td>
<td>42</td>
<td>696</td>
<td>1592</td>
<td>1592</td>
<td>1372</td>
<td>1140</td>
<td>1019</td>
<td>1342</td>
<td>1576</td>
</tr>
<tr>
<td>MTPE</td>
<td>131</td>
<td>1890</td>
<td>1804</td>
<td>1743</td>
<td>1369</td>
<td>1141</td>
<td>922</td>
<td>1481</td>
<td>1433</td>
</tr>
<tr>
<td>Trans.</td>
<td>132</td>
<td>1914</td>
<td>1701</td>
<td>1319</td>
<td>993</td>
<td>916</td>
<td>933</td>
<td>1236</td>
<td>1223</td>
</tr>
<tr>
<td>MTPE gain%</td>
<td>-</td>
<td>-</td>
<td>6.06</td>
<td>32.15</td>
<td>37.78</td>
<td>24.62</td>
<td>-1.23</td>
<td>19.80</td>
<td>17.20</td>
</tr>
</tbody>
</table>

Table 3: Productivity achieved per translator and match band in words per hour.

The throughputs obtained seem a bit high when compared to the usual reference values of 313–375 words per hour (2500–3000 words per day) for translation tasks. While it is true that the content we usually receive from this client often allows our translators to translate faster than their regular throughput, the values reported here are still exceptionally high. To explain them, it must be noted that the productivity in Table 3 only reflects the time spent by translators with the cursor placed inside a segment in memoQ’s editor. Other tasks such as reading of instructions, file management, escalation of queries, self-revision, terminological investigation, quality assurance and communication with project managers and/or other parties involved in the project are not reflected here, but are all part of a regular project. Care should also be taken when transferring productivity gains into rate discounts, as translation rates often include tasks which are not affected by the introduction of MTPE (such as project management, file engineering or revision by a second linguist).

Examples 2–4 are samples of text to be translated in the project. As may be observed, the text is written in plain English and the sentences are not too complex.

(2) {1}Activate trial version of the application{2}.
(3) When the trial license expires, the trial version of the application cannot be activated for a second time.
(4) This check box enables / disables the option that prevents Firewall from stopping until the operating system shuts down completely.

As shown in the last row of Table 3, all translators except one (Translator 6, who experienced 1.2% productivity loss when facing MTPE) were faster in MTPE than translating from scratch. However, the productivity gains have a great variability ranging from 6.06% to 56.34%. Other researchers studying MTPE have also reported this great variation in their experiments.
(Guerberof Arenas, 2009; Plitt and Masselot, 2010; Koehn and Germann, 2014). For instance, productivity gains between 20% and 131% were reported by Plitt and Masselot (2010).

A possible explanation for the slight productivity loss experienced by Translator 6 might be that this translator had experienced an extensive period of inactivity and had barely used memoQ. Also, the biggest productivity gain was achieved by the least experienced translator (Translator 10), while the smallest productivity gain corresponds to the most experienced one (Translator 1). However, this trend cannot be confirmed by the rest of results.\(^7\)

### 3.2 MT evaluation metrics

Since collecting and comparing translation-from-scratch and MTPE throughputs in each project is time-consuming and may not always be feasible, it is common practice to use automated metrics as an indicator of productivity gains. We calculated document-level and segment-level values for BLEU,\(^8\) TER and target-side FMS taking the segments belonging to each band as separate documents. BLEU and TER were obtained using Asiya (Giménez and Márquez, 2010), whereas the FMS was computed using Okapi framework’s Rainbow application via its Translation Comparison feature.\(^9\) This target-side FMS is based on the Sørensen-Dice coefficient (Sørensen, 1948; Dice, 1945) using character 3-grams. One difference with both BLEU and TER is that this FMS is applied to character rather than word sequences.

Table 4 reports the average results for each metric and the average gain obtained. In each segment, we computed the automatic metrics and productivity gain for each translator individually using their specific throughputs and text output. This means that in no case we calculated the metrics using multiple references. The gain is calculated according to Equation 1 above.

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>TER</th>
<th>FMS</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>92.68</td>
<td>4.10</td>
<td>97.48</td>
<td>127.38%</td>
</tr>
<tr>
<td>95-99%</td>
<td>85.35</td>
<td>9.19</td>
<td>92.12</td>
<td>66.20%</td>
</tr>
<tr>
<td>85-94%</td>
<td>82.31</td>
<td>11.99</td>
<td>91.38</td>
<td>76.82%</td>
</tr>
<tr>
<td>75-84%</td>
<td>70.70</td>
<td>20.98</td>
<td>85.60</td>
<td>22.52%</td>
</tr>
<tr>
<td>MTPE</td>
<td>66.07</td>
<td>20.90</td>
<td>87.91</td>
<td>24.09%</td>
</tr>
</tbody>
</table>

Table 4: Automatic scores and productivity gain for each band.

A surprising finding is the average gain reported for the 95–99% band. It would be logical to expect the gain of this band being somewhere in between the 100% and the 85–94% band values, as the three automatic metrics actually indicate, but it turned out to be inferior to the 85–94% band. This can be explained by the fact that the vast majority of edits required in the 95–99% band involved dealing exclusively with inline tags. Although the impact of these operations has not been researched enough, these results show that they can have a big impact in terms of productivity, slowing down the translator more than it would be expected.

Moreover, when calculating automated metrics, inline tags are generally deleted to avoid their division in different tokens. Instead of deleting them, when calculating the automated metrics reported in this paper we converted each tag into a unique token. Although this operation brought the 95–99% values closer to the 85–94% band, it was not enough to compensate all the effort put into tag handling, as hinted by the productivity gain values. More research is needed

\(^7\)For further discussion on the impact of experience in the throughputs reported in the experiment, see Parra Escartín and Arcedillo (2015b).
\(^8\)As stated in Asiya’s documentation, BLEU scores are estimated using the NIST’s script used for evaluation campaigns and available at: \(\text{http://www.itl.nist.gov/iad/mig//tools/}\).
\(^9\)We used Rainbow (\(\text{http://okapi.opentag.com}\)) because it natively supports the most common bilingual formats in the industry, already computes a target-side FMS and is freely available, thus potentially improving transparency and usability of MTPE evaluation.
in this area to find out the appropriate weight or penalty that automated metrics should assign to inline tags. For this reason, we opted to treat the 95–99% band as an outlier and ignore it from further analysis.

4 Productivity gain thresholds

With the aim of establishing where the productivity gain threshold lies, we crossed productivity values with automatic MT evaluation metrics. For each translator in the experiment, we estimated the sentence-level BLEU, TER and FMS values in the MTPE and TM match samples.

Figures 1, 3 and 5 show the overall results for each evaluation metric, while Figures 2, 4 and 6 show the number of segments for each band and metric. The gain values can be expected to be more reliable the higher the number of segments which fell in that band.

As can be observed in Figure 1, the last BLEU band which reported productivity gains is the 45–50 band. Between 35 and 45, the productivity is slightly inferior to the translation-from-scratch average. Below 35, the productivity decreases more drastically. This finding is interesting, as it is normally said that a BLEU score of 30 reflects understandable translations, while scores over 50 can be considered good and fluent translations (Lavie, 2010). Even though a BLEU score of 30–45 may be useful for other MT applications (such as gisting), in the case of MTPE it does not seem to yield any productivity increase.

In our experience in past projects, BLEU values above 50 are a clear indicator that productivity gains can be achieved, while gains in the 45–50 range are common, but cannot be taken for granted. Below 45, we have never experienced productivity gains, so the findings reported here seem to support our past experiences. These experiences cover a broad range of domains, all types of MT systems (generic, customized, rule-based, statistical, hybrid, etc.) and diverse quality of MT output. However, they only encompass English into Spanish tasks and different results are likely to be experienced by other language pairs.

As far as the productivity gains for each band are concerned, it is also noticeable that the BLEU 90–95 band obtains a greater productivity gain than the 100 mark. The few number of segments within this band (11), significantly lower than segments in the 100 and the 85–90 groups, may have skewed the results for this particular band. Even though there are other examples of inconsistencies with contiguous bands (such as the productivity of the 60–65 band being a bit lower than expected and the one for 75–80 being a bit higher), the trend seems clear and our results point to a productivity gain threshold around 45 BLEU score. Apart of the few number of segments in the 90–95 range, it is also noticeable that no segment fell into the 95–99 BLEU band, despite the relatively large number of segments in the 100 and 85–90 ranges.
The TER results are more difficult to interpret (Figure 3). The productivity gain drops slightly below translation-from-scratch average at the 30–35 band, rises above this average at 35–40 and drops again for good at the 40–45 mark. According to this data, the last TER range with clear productivity increase would be the 25–30 band, with the tipping point somewhere between 30 and 40. In our past experiences, the threshold for productivity increase was also situated around 35, although its variability proved to be higher than BLEU’s.

![Figure 3: TER scores vs. productivity gain.](image)

The FMS, on the other hand, shows a more consistent monotonic trend (Figure 5). Also, the distribution of the number of segments within the different bands is closer to a standard normal distribution than the other metrics evaluated (Figure 6). The last band where productivity gain is reported is the 75–80 range. The productivity for 70–75 falls slightly below translation-from-scratch average, while below 70 the productivity loss is more dramatic. Therefore, the productivity gain threshold for FMS seems to lie at the 75% mark. It is interesting that these target-side FMS results match the traditional industry practice of not offering discounts for TM matches below a source-side FMS of 75%, following the general assumption that they do not yield any productivity increase.

It is tempting to take this analogy further and try to apply the TM match pricing schemes to MTPE tasks (i.e., apply to unedited MT segments the same discount as 100% TM matches, etc.). In the next section we compare throughputs of TM matches and MTPE segments in order to evaluate the appropriateness of this approach.

![Figure 5: FMS vs. productivity gain.](image)

![Figure 6: Segments per FMS band.](image)
5 Productivity and cognitive effort in TM matches and MTPE segments

Since in our experiment a significant amount of TM fuzzy matches were post-edited alongside MTPE segments, exact matches and no-match segments, we can compare the productivity of segments which required equal amount of text editing but belong to different categories. For example, we can compare the productivity of MTPE segments which required no editing with unmodified 100% TM matches, or MTPE segments with 75–84% target-side FMS with its equivalent segments from the TM sample.\textsuperscript{10} Our hypothesis is that MTPE segments are slower to post-edit because they require a higher cognitive effort\textsuperscript{11} to identify the parts of the sentence that need (or do not need) editing, while when post-editing TM matches those parts are already highlighted for the user by the CAT tool.

Another factor is that translators may consider TM suggestions more reliable than MT output. In this experiment, the TMs provided to the translators were the actual ones used in production with this particular client. All segments contained in these TMs have been self-revised by the translator, revised by a second linguist and validated by the client at least once over the past few years. Therefore, the matches retrieved can be considered reliable and representative of the quality expected by the client. If this same experiment is re-run using less reliable TMs (for example, ones with no terminological, syntactical or orthotipographic consistency), it is possible that the effort put into achieving structural and terminological homogeneity, or even correcting translation errors, would negate part or all the productivity gains reported below of TM matches over high-quality MT output.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure7.png}
\caption{Productivity of unedited 100% matches and unedited MTPE segments.}
\end{figure}

Figure 7 shows the throughput of segments which were left unedited from the 100% TM match and MTPE bands. All translators except one worked faster with these unedited 100% matches than with the MTPE sub-sample. The exception is Translator 10, the least experienced translator, whose productivity in both categories was almost the same. Since in both categories the amount of editing was the same (in this case, no edits were performed), the difference may lie in the higher cognitive effort invested in the MTPE segments, or that the translator had less confidence in the MT suggestion and spent more time checking it.

To study if this behaviour replicates when text editing is involved, we compared segments

\textsuperscript{10}It should be noted that different CAT tools use different implementations of the FMS. Thus, the CAT tool used may also have an impact on the results obtained. Replicating the experiment reported here using different CAT tools would be need to shed light on this issue.

\textsuperscript{11}Previous studies (Koponen et al. (2012)) have already successfully applied post-editing time as a way to assess the cognitive effort involved in MTPE.
with similar FMS values both from the MTPE and 75–84% TM match sets. We selected that TM match set because it is the one closest to the MTPE sample in terms of number of words per segment, productivity and automatic scores (Tables 1, 3 and 4 respectively). We then selected only the segments from those bands with a target-side FMS of 75–84 to be sure that both sets involved the same amount of target text editing. Figure 8 shows the productivity obtained by each translator in both sets. Again, all translators except the least experienced one (Translator 10) worked slower with the MTPE sub-sample than when post-editing TM matches.

![Figure 8: Productivity of 75–84% TM matches and MTPE segments with 75–84 FMS.](image)

These results indicate that, even though the edits performed in both samples were the same, when facing MTPE samples translators need to invest more cognitive effort deciding which parts of the sentence need correction. When leveraging TM segments, however, these parts are already highlighted by the CAT tool, thus making them faster to post-edit. It may be worth researching a similar feature for MTPE segments, where the parts of the sentence with lowest confidence are highlighted for the user.\footnote{Nayek et al. (2015) have already considered this approach in the context of the development of a post-editing tool.}

One of the conclusions that can be drawn from these results is that the pricing system of source-side FMS due to TM leverage cannot be directly applied to MTPE. The different productivity gains reported for these two kinds of translation technologies indicate that MT segments are slower to post-edit than equivalent TM matches, even where no editing is involved. Therefore, it would not be adequate to apply the 100% match rate discount to MTPE segments that require no editing, or the low fuzzy match rate discount to equivalent MTPE segments.

We are aware that other researchers have reported equivalences between MTPE and TM match editing in the past. For example, Guerberof Arenas (2009) concludes that MTPE is equal to 85–94% TM fuzzy editing. However, in these experiments TM matches were not used in their usual settings, as the translator did not know if the suggestion came from MT or TM, and therefore no highlighting of the text differences was offered to the translator. For this same reason, the conclusion that the quality of post-edited output from MT is higher than post-edited TM output cannot be applied to a real commercial environment as the one we tried to replicate here.

### 6 Conclusion and future work

In this paper, we reported an experiment where ten professional translators with diverse experience in translation and MTPE complete the same translation and post-editing task within their everyday work environment using files from a real translation request. The more than 7,000...
words of the file to be translated included a significant amount of TM fuzzy matches, TM exact matches and no-match segments. Half of the no-match segments was randomly selected for MTPE, while the other half was translated from scratch. The MT output for the MTPE sample was generated using one of our customized RBMT engines.

We compared the productivity gain achieved due to MTPE with automated metrics such as BLEU, TER and FMS in order to find out the threshold at which each metric starts to consider that productivity gains can be achieved. According to our results, BLEU scores above 45, TER values below 30 and FMS values above 75 mean that productivity increases can be obtained. We have also detected a grey area in TER values between 30 and 40, in which the productivity increase is difficult to interpret. These thresholds agree with past experiences in our company, although only the pair English into Spanish has been considered and different performance by other pairs is to be expected.

A comparison between equivalent segments from the MTPE and TM matching samples has shown that, where equivalent text editing is involved, MTPE segments are slower to post-edit than TM matches, even where the TM/MT output was left unmodified. This can be explained by the higher cognitive effort required in MTPE segments, where the translator needs to identify the parts of the sentence which need editing, as opposed to having the CAT tool highlight these parts automatically in TM leveraging. Translators’ confidence in TM suggestion may also play a role here. A side effect of these findings is that MT output cannot be assigned the same pricing system as TM fuzzy matching, although using fuzzy scores for MTPE evaluation may help to map these two systems and create intuitive analogies.

As future work, it may be interesting to find out if highlighting the parts of the sentence which the MT system feels less confident about would reduce the cognitive effort involved in MTPE. We also plan to apply quality estimation techniques to this same data set in order to prevent poor MT output from being post-edited, thus eliminating cases of productivity loss when compared to translation from scratch.

Acknowledgments

The research reported in this paper is supported by the People Programme (Marie Curie Actions) of the European Union’s Framework Programme (FP7/2007-2013) under REA grant agreement n° 317471.

We would like to thank our colleagues at Hermes: The project managers, for allocating the time and resources that allowed us to run the experiment, and the translators who participated in it and provided us with the data analysed in this paper. We would also like to thank the anonymous reviewers for their valuable feedback to improve this paper and for the ideas for future work they have provided us with.

References


Domain Adaptation for Social Localisation-based SMT: A Case Study Using the Trommons Platform

Jinhua Du jdu@computing.dcu.ie
Andy Way away@computing.dcu.ie
Zhengwei Qiu zhengwei.qiu@computing.dcu.ie
ADAPT Centre, School of Computing, Dublin City University, Ireland
Asanka Wasala Asanka.Wasala@ul.ie
Reinhard Schaler reinhard.schaler@ul.ie
Localisation Research Centre, University of Limerick, Ireland

Abstract

Social localisation is a kind of community action, which matches communities and the content they need, and supports their localisation efforts. The goal of social localisation-based statistical machine translation (SL-SMT) is to support and bridge global communities exchanging any type of digital content across different languages and cultures. Trommons is an open platform maintained by The Rosetta Foundation to connect non-profit translation projects and organisations with the skills and interests of volunteer translators, where they can translate, post-edit or proofread different types of documents. Using Trommons as the experimental platform, this paper focuses on domain adaptation techniques to augment SL-SMT to facilitate translators/post-editors. Specifically, the Cross Entropy Difference algorithm is used to adapt Europarl data to the social localisation data. Experimental results on English–Spanish show that the domain adaptation techniques can significantly improve translation performance by 6.82 absolute BLEU points and 5.99 absolute TER points compared to the baseline.

1 Introduction

The concept of social localisation was proposed In Schäler (2011), which is based on the fact that currently large communities with language skills are ready to support good causes, and large amounts of content are accessible to communities that should be translated, but is not being done currently. Accordingly, social localisation aims to match communities and slice content, and support the localisation efforts.¹

The main objective of social localisation is the promotion of a demand– rather than a supply–driven approach to localisation. It is based on the recognition that it is no longer exclusively the fact that the corporations control the global conversation, but rather communities. Social localisation supports user-driven and needs-based localisation scenarios, in contrast to mainstream localisation, driven primarily by short-term financial return-on-investment considerations.

Trommons² - short for Translation Commons - is an open platform maintained by The Rosetta Foundation³ to connect non-profit translation projects and organisations with the skills

¹https://en.wikipedia.org/wiki/Social_localisation
²http://trommons.org/
³http://www.therosettafoundation.org/
and interests of volunteer translators. The idea behind Trommons is to match translation/post-editors/proofreading tasks published by various NGOs with volunteer translators/users registered in Trommons.

It is well-known that SMT systems have been widely deployed into the translator’s workflow in the localisation and translation industry to improve productivity, which is also named as post-editing SMT (PE-SMT) (Guerberof, 2009; Plitt and Masselot, 2010; Carl et al., 2011; O’Brien, 2011; Zhechev, 2012; Guerberof, 2013). In Trommons, many language pairs lack high-level translators, or translation jobs take quite a long time to be claimed and finished, so a PE-SMT system for such language pairs or translation projects are a practical solution in which human effort could be significantly reduced and high-quality translations could be achieved.

However, data is a big problem for building high-quality PE-SMT systems in Trommons. In this paper, we use domain adaptation techniques to acquire social localisation domain related data to augment PE-SMT systems. Specifically, a Cross Entropy Difference (CED) method is used to select different scales of data from the Europarl data sets. Experiments conducted on English-to-Spanish show that the domain adaptation method can improve translation quality by 6.82 absolute (20.98 relative) BLEU points (Papineni et al., 2002) and 5.99 absolute (11.26 relative) TER points (Snover et al., 2006) compared to the baseline.

2 Related Work

Domain adaptation is a popular but difficult research question in SMT. It is well-known that SMT performance is heavily dependent on the training data and the development set (devset) as well as the estimated model parameters which can best reflect the characteristics of the training data. Therefore, if the characteristics of the test data are substantially different from those of the model parameters, system performance drops significantly. In this case, the domain adaptation can be used to adapt an SMT system to the out-of-domain (OOD) test data in order to improve translation performance.

For social localisation data in Trommons, the currently available in-domain (ID) data for each language pair is far from sufficient to build up a high-quality SMT system (cf. Section 5.2). Therefore, we have to use a data selection algorithm to select ID data from the OOD data in order to augment the SL-SMT.

Lv et al. (2007) use information-retrieval techniques in a transductive learning framework to increase the count of important ID training instances, which results in phrase-pair weights being favourable to the devset. Bicici and Yuret (2011) employ a Feature Decay Algorithm (FDA) to increase the variety of the training set by devaluing features that have already been seen from a training set, and experimental results show significant improvements compared to the baseline.

There has been increasing interest in applying the CED method to the problem of SMT data selection (Moore and Lewis, 2010; Axelrod et al., 2011; Haque et al., 2014). In this method, given an ID corpus $I$ and a general corpus $O$, language models are built from both, and each sentence $s$ in $O$ is scored according to the entropy difference, as in Equation (1):

$$\text{score}(s) = H_I(s) - H_O(s)$$  \hspace{1cm} (1)

where $H_I(s)$ is the entropy of $s$ in ID corpus $I$, and $H_O(s)$ is the entropy of $s$ in OOD corpus $O$. The score$(s)$ is to reflect how similar the sentence $s$ is to the corpus $I$, and how different the sentence $s$ is from the corpus $O$. That is, the lower the score given to a sentence, the more useful it is to train a system for the specific domain $I$.

Some other methods have been proposed to select ID data. For example, Banerjee et al. (2012) propose an approach to perform batch selection with the objective of maximizing SMT performance. Toral (2013) and Toral et al. (2014) use linguistic information such as lemmas,
named entity categories and part-of-speech tags to augment perplexity-based data selection and achieved improved results.

The CED method is more promising both in domain adaptation for SMT and data selection for active learning-based PE-SMT (Dara et al., 2014), so we choose it to verify its effectiveness in SL-SMT application.

3 System Description of Trommons

In Trommons, registered organisations/NGOs can create projects. Projects are identified by project IDs. A project can have a source document and a number of tasks associated with it. These tasks can be one of the following types: translation, proofreading or segmentation. All the files that belong to a certain project are stored in the project folder.

The system stores raw files in a MySQL database. The files can be in different formats (e.g. pdf, doc, xliff, ppt). Multiple versions of the files are stored in the file system, i.e. version 0 is normally the source file (v-0 folder), subsequent versions may denote the translation or proof-read version of the source file (e.g. v-1, v-2 folders). In the MySQL database, Project table contains the necessary information about the project; Tasks table includes different information about the tasks, including the language pairs, countries, organisations etc. Different task types are defined in the TaskTypes table, e.g., 3 indicates “Proofreading” and 2 indicates “Translation”. The tasks can be in different states (claimed, completed etc.) which are defined in TaskStatus table. Using these tables a query can be used to retrieve the information, e.g. an organisation can be retrieved from the Organisations table.

Some projects may have been already completed, and the system allows archiving of such old projects. Data related to these projects can be retrieved from the Archived Projects table and the Archived Tasks table. Their associated metadata can be found in the ArchivedProjectsMetadata table. The TaskFileVersions table includes details of the latest version of the files (i.e. v-0, v-1 etc in the file system).

However, in Trommons, the parallel texts are not automatically maintained in the file system, so we have to extract and align source and target sentences from the raw files. In addition, information about the language pairs is not maintained in the file system; it can only be acquired from the database.

4 Data Statistics and Analysis

We export data from the current Trommons database, and examine the tables and data according to descriptions of the file system.

4.1 Data Statistics

The data statistics are listed in Table 1.

<table>
<thead>
<tr>
<th>#Projects</th>
<th>#Tasks</th>
<th>#Organisations</th>
<th>#Languages</th>
<th>#Countries</th>
<th>#FileTypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,270</td>
<td>7,825</td>
<td>263</td>
<td>7,431</td>
<td>250</td>
<td>xls, doc, docx, zip, pdf, odt, dot, ppt, po, xlsx, xliff</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the Trommons database

In Table 1, ‘#Projects’ indicates the number of projects in the database, ‘#Tasks’ is the total number in all projects, ‘#Organisations’ is the number of different ‘Organisations’ registered in Trommons, ‘#Languages’ indicates the number of different languages (including dialects, sign language etc.) that are used in projects, ‘#Countries’ is the number of different countries where translators or organisations come from, and ‘#FileTypes’ indicates different types of documents stored in the database.
From the data statistics, we can see that Trommons has successfully connected non-profit translation projects and organisations with thousands of volunteer translators worldwide.

4.2 Data Analysis

We query the ‘completed translation’ tasks from the database, and retrieve 1,990 tasks that belong to 964 projects, containing 23 languages in terms of the source language and 63 languages in terms of the target language, coming from 44 countries regarding the source language and 74 countries regarding the target language. In these 1,990 tasks, there are 101 language pairs in total.

The breakdown of TOP-10 ranks in terms of language pairs, source language, target language and countries are shown in Table 2.

<table>
<thead>
<tr>
<th>No.</th>
<th>source language</th>
<th>target language</th>
<th>#tasks</th>
<th>ratio(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English</td>
<td>Spanish</td>
<td>422</td>
<td>21.21</td>
</tr>
<tr>
<td>2</td>
<td>Spanish</td>
<td>English</td>
<td>266</td>
<td>13.37</td>
</tr>
<tr>
<td>3</td>
<td>English</td>
<td>French</td>
<td>152</td>
<td>7.64</td>
</tr>
<tr>
<td>4</td>
<td>French</td>
<td>English</td>
<td>105</td>
<td>5.28</td>
</tr>
<tr>
<td>5</td>
<td>English</td>
<td>Portuguese</td>
<td>94</td>
<td>4.72</td>
</tr>
<tr>
<td>6</td>
<td>English</td>
<td>Russian</td>
<td>87</td>
<td>4.37</td>
</tr>
<tr>
<td>7</td>
<td>English</td>
<td>Arabic</td>
<td>76</td>
<td>3.82</td>
</tr>
<tr>
<td>8</td>
<td>English</td>
<td>Chinese</td>
<td>71</td>
<td>3.57</td>
</tr>
<tr>
<td>9</td>
<td>Spanish</td>
<td>French</td>
<td>65</td>
<td>3.27</td>
</tr>
<tr>
<td>10</td>
<td>French</td>
<td>Spanish</td>
<td>59</td>
<td>2.96</td>
</tr>
</tbody>
</table>

Table 2: Top-10 language pairs in terms of tasks

In Table 2, we can see that 1) English-based language pairs have the largest proportion in terms of the source language; 2) Spanish-based language pairs have the largest proportion in terms of the target language; 3) the tasks of the top-10 language pairs account for 70.2% in all 1,990 tasks.

<table>
<thead>
<tr>
<th>No.</th>
<th>source language-based tasks</th>
<th>target language-based tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>language</td>
<td>#tasks</td>
</tr>
<tr>
<td>1</td>
<td>English</td>
<td>1,280</td>
</tr>
<tr>
<td>2</td>
<td>Spanish</td>
<td>389</td>
</tr>
<tr>
<td>3</td>
<td>French</td>
<td>167</td>
</tr>
<tr>
<td>4</td>
<td>Portuguese</td>
<td>29</td>
</tr>
<tr>
<td>5</td>
<td>German</td>
<td>27</td>
</tr>
<tr>
<td>6</td>
<td>Catalan</td>
<td>24</td>
</tr>
<tr>
<td>7</td>
<td>Italian</td>
<td>18</td>
</tr>
<tr>
<td>8</td>
<td>Hebrew</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>Chinese</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>Swedish</td>
<td>7</td>
</tr>
<tr>
<td>total</td>
<td>–</td>
<td>1,960</td>
</tr>
</tbody>
</table>

Table 3: Top-10 languages in terms of source side and target side, respectively

In Table 3, we can see that 1) English-based tasks account for the largest proportion in terms of the source-side language; 2) English-based and Spanish-based tasks account for the
same proportion that is much greater than others in terms of the target language; 3) the tasks of the top-10 source-side languages and target-side languages account for 98.49% and 85.28%, respectively.

<table>
<thead>
<tr>
<th>No.</th>
<th>countries based on source-side language</th>
<th>countries based on target-side language</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>country</td>
<td>#tasks</td>
</tr>
<tr>
<td>1</td>
<td>United States</td>
<td>540</td>
</tr>
<tr>
<td>2</td>
<td>United Kingdom</td>
<td>474</td>
</tr>
<tr>
<td>3</td>
<td>Spain</td>
<td>195</td>
</tr>
<tr>
<td>4</td>
<td>Ireland</td>
<td>156</td>
</tr>
<tr>
<td>5</td>
<td>ANY</td>
<td>148</td>
</tr>
<tr>
<td>6</td>
<td>France</td>
<td>139</td>
</tr>
<tr>
<td>7</td>
<td>Colombia</td>
<td>67</td>
</tr>
<tr>
<td>8</td>
<td>Mexico</td>
<td>52</td>
</tr>
<tr>
<td>9</td>
<td>Brazil</td>
<td>29</td>
</tr>
<tr>
<td>10</td>
<td>Congo</td>
<td>27</td>
</tr>
<tr>
<td>total</td>
<td>–</td>
<td>1,827</td>
</tr>
</tbody>
</table>

Table 4: Top 10 countries in terms of source-side and target-side languages, respectively

In Table 4, ‘ANY’ indicates the country that was not specified when creating a task. We can see that 1) United States has the largest proportion of translation tasks in terms of both source-side language and target-side language; 2) Other countries such as United Kingdom, Spain, France and Ireland have a large proportion compared to the rest of the countries.

We can conclude from the statistics that in Trommons:

- English–Spanish is the most popular language pair, which also indicates that the number of translators of this language pair is the largest;
- English-oriented and Spanish-oriented tasks account for larger proportions than other languages;
- The United States, United Kingdom, Spain and France account for most of the tasks, but countries like Colombia, Mexico, Brazil, Peru account for big proportions as well because these countries use Spanish and Portuguese.

5 Experiment Preparation

Having examined what data is available, we now report on two experiments we carried out, namely the ID SMT experiment and domain adaptation SMT experiment. In the latter, the CED data selection method is used to acquire more ID data from the OOD data. The Experimental details are described in the following sections.

5.1 Experiment Plans and Considerations

We carried out the following steps in our experiments:

- Examination of the tasks that are “translation” type and which have been “completed”, and recognition of how many language pairs and file types can be retrieved.
- Extraction of parallel data from the task files above.
- Sentence alignment and creating and evaluating sentence-level parallel data, and evaluate the parallel data.
• Creation of development set and test set.
• Development of SMT system and tuning of parameters.
• Evaluation of the SMT system.
• Selection of data from Europarl using CED method and building of domain-adaptive SL-SMT, and evaluation of its performance.

5.2 Data Pre-processing and Sentence Alignment

In Table 2, the English–Spanish pair has the biggest proportion of tasks among all language pairs, so we selected this for SMT system creation.

262 pairs of documents we retrieved for the English–Spanish pair, in which 231 documents were `.doc` or `.docx` files. These types of documents needed to be converted to `.txt` files for SMT system building.

We used the ‘Hunalign’ toolkit (Varga et al., 2005) to run the sentence alignment, and obtained 10,151 aligned sentence pairs. After filtering out the longer (>100) sentences, we ended up with 9,379 parallel sentences.

We then randomly selected out 100 sentence pairs in all 9,379 parallel sentences to evaluate the accuracy of aligned sentences by human, and we achieve the accuracy of 98%.

6 SMT Experiments

Due to the insufficient amount of parallel data extracted from the Trommons tasks, domain adaptation techniques had to be used to acquire more ID parallel data to obtain reasonable translation performance.

In order to build the SMT system, we use Moses (Koehn et al., 2007) with default settings. For the data sets, we randomly extracted two sets of 500 sentence pairs from the 9,379 pairs of sentences as the devset and test set, respectively. The remaining 8,379 pairs of parallel data was used to build the initial ID SL-SMT system.

We used the CED method to select different scales of ID data from the Europarl data set. The statistics of the different adapted data sets are shown in Table 5.

<table>
<thead>
<tr>
<th>system</th>
<th>#sentence</th>
<th>#word</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>Spanish</td>
</tr>
<tr>
<td>baseline0</td>
<td>8,379</td>
<td>171,593</td>
</tr>
<tr>
<td>baseline1</td>
<td>3,993,529</td>
<td>104,280,921</td>
</tr>
<tr>
<td>baseline2</td>
<td>3,993,529</td>
<td>104,280,921</td>
</tr>
<tr>
<td>DaM1</td>
<td>120,566</td>
<td>2,671,412</td>
</tr>
<tr>
<td>DaM2</td>
<td>205,593</td>
<td>5,171,210</td>
</tr>
<tr>
<td>DaM3</td>
<td>284,370</td>
<td>7,671,080</td>
</tr>
<tr>
<td>DaM4</td>
<td>363,103</td>
<td>10,170,859</td>
</tr>
<tr>
<td>DaM5</td>
<td>827,362</td>
<td>25,169,906</td>
</tr>
<tr>
<td>DaM6</td>
<td>982,405</td>
<td>30,169,652</td>
</tr>
<tr>
<td>DaM7</td>
<td>1,138,302</td>
<td>35,169,395</td>
</tr>
<tr>
<td>DaM8</td>
<td>1,295,919</td>
<td>40,169,132</td>
</tr>
<tr>
<td>DaM9</td>
<td>1,456,383</td>
<td>45,168,909</td>
</tr>
<tr>
<td>DaM10</td>
<td>1,620,025</td>
<td>50,168,648</td>
</tr>
</tbody>
</table>

Table 5: Statistics of different data sets
In Table 5, “baseline0” indicates that the data is the extracted parallel data from the Trommons platform. “baseline1” and “baseline2” use 3.9 million pairs of Europarl data, in which the difference is that “baseline1” is tuned on the WMT newswire2012 testset, while “baseline2” is tuned on the extracted devset. “DaM1” indicates a combination of Trommons-extracted data and the adaptive data extracted from the Europarl data, and the same for “DaM2”~“DaM10”. The difference in these adaptive data is the amount of data used.

The results of the baselines and adaptive SMT systems are shown in Table 6.

<table>
<thead>
<tr>
<th>system</th>
<th>BLEU4</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline0</td>
<td>32.50</td>
<td>53.20</td>
</tr>
<tr>
<td>baseline1</td>
<td>30.98</td>
<td>55.83</td>
</tr>
<tr>
<td>baseline2</td>
<td>32.57</td>
<td>52.78</td>
</tr>
<tr>
<td>DaM1</td>
<td>37.89</td>
<td>48.72</td>
</tr>
<tr>
<td>DaM2</td>
<td>37.89</td>
<td>48.60</td>
</tr>
<tr>
<td>DaM3</td>
<td>38.56</td>
<td>48.44</td>
</tr>
<tr>
<td>DaM4</td>
<td>38.41</td>
<td>48.11</td>
</tr>
<tr>
<td>DaM5</td>
<td>39.19</td>
<td>47.21</td>
</tr>
<tr>
<td>DaM6</td>
<td>38.60</td>
<td>48.32</td>
</tr>
<tr>
<td>DaM7</td>
<td><strong>39.32</strong></td>
<td>47.45</td>
</tr>
<tr>
<td>DaM8</td>
<td>38.91</td>
<td>47.86</td>
</tr>
<tr>
<td>DaM9</td>
<td>38.61</td>
<td>48.22</td>
</tr>
<tr>
<td>DaM10</td>
<td>38.78</td>
<td>47.92</td>
</tr>
</tbody>
</table>

Table 6: Results of different SMT systems

In Table 6, we can see that:

- “baseline0”, “baseline1” and “baseline2” have similar performance, but the data sizes of “baseline1” and “baseline2” are far larger than that of “baseline0”, which indicates that there is a significant difference between the Europarl data and the social localisation data in terms of domain.

- The maximum improvement is 6.82 absolute (20.98 relative) BLEU points and is 5.99 absolute (11.26 relative) TER points compared to the “baseline0”.

- Immediately, with the increase in training data size selected from the Europarl data using our domain adaptation technique, translation performance correspondingly increases, e.g. the BLEU score is 32.50 for “baseline0”, and 37.89 for DaM1 and 39.32 for DaM7, where we see the improvements by 5.39 and 6.82 absolute BLEU points. It is almost the same trend for TER score, where we see the improvements by 4.48 and 5.75 absolute TER points.

- However, when the amount of the data increases to some scale, e.g. from DaM4, translation performance becomes unstable, with performance decreases in some cases, despite the fact that the amount of the data increases.

- Data selection can effectively select appropriate data and keep the training data to a reasonable scale. More importantly, it can be used to quickly build and deploy a relatively high-quality SMT system for practical use.
7 Conclusions

In this paper, we carried out a case study using domain adaptation techniques to augment an SL-SMT system applied in Trommons platform for the English–Spanish language pair. We first examined the file systems in the Trommon platform, and transformed Microsoft Office documents to plain text to be used for SMT building, and then extracted parallel sentences with the alignment rate of 98%. Finally, we used the CED method to select different sizes of data from the Europarl corpus to augment the initial SL-SMT system and achieved improvements of maximum 6.82 absolute BLEU points and 5.99 absolute TER points compared to the baseline.

In future work, we intend to carry out further studies on SL-SMT regarding 1) online domain adaptation: for a more practical SL-SMT system, online domain adaptation can incrementally select ID data to augment SMT system; 2) developing a more effective data selection algorithm to select ID data and localized data; 3) applying our method to new language pairs and domains.

Acknowledgments

We thank the reviewers for their insightful comments. This research is supported by Science Foundation Ireland through the ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie) at Dublin City University, and by Grant 610879 for the Falcon project funded by the European Commission.

References


