On the correctness of machine translation: A machine translation post-editing task

Maarit Koponen, University of Helsinki, and Leena Salmi, University of Turku, Finland

ABSTRACT

Machine translated texts are increasingly used for quickly obtaining an idea of the content of a text and as a basis for editing the text for publication. This paper presents a study examining how well a machine-translated text can convey the intended meaning to the reader.

In the experiment described, test subjects edited machine-translated texts from English into Finnish. In order to observe how well it would be possible for the test subjects to decipher the meaning of the source text based on the machine translation alone, they had no access to the source text. Their edits were assessed by the authors of the paper for the correctness of meaning (compared to the source text) and language (compared to the target language norms and conventions). The results show that the test subjects were successful at deducing the correct meaning without the source text for about half of the edited sentences. The results also suggest that errors in word forms and mangled relations that can be deduced based on context are the kind of machine translation errors that are easier to recover from, while mistranslated idioms and missing content seem to be more critical to understand the meaning.

KEYWORDS

Machine translation, post-editing, translation quality, evaluation, translation quality assessment.

1. Introduction

The growing amount of material to be translated, the pressure to produce translations faster, and efforts to lower costs have led to renewed interest in the development of machine translation (MT) systems. Growing interest and recent advances particularly in statistical approaches are making MT increasingly common in both professional translation contexts and everyday life. Many professional translators' work now involves the use of machine-translated texts for post-editing – according to Robert (2013: 32), it can increase the average number of words translated by a professional from 2000 to 3500 words per day. At the same time, the various translation systems available online, often for free, are increasingly used for purposes such as 'gisting' or quickly forming a basic idea of the content of a text. This interest in and use of MT also means that there is a growing need for the Translation Studies community to actively contribute to its development, as noted by Rebecca Fiederer and Sharon O'Brien (2009: 52). Much of the research into MT and post-editing processes in recent years, for example by Plitt and Masselot (2010) and Carl *et al.* (2011), has focused on productivity: editing speed compared to translation from scratch, amount of editing performed and the quality achievable through post-editing.

One area where translators and Translation Studies may have valuable contributions is translation quality evaluation from the point of view of the MT user and the quality criteria set by the purpose of translation. As Lori Thicke (2013: 10) has noted, in the increasingly common scenario of MT post-editing workflows, the post-editors are ultimately the ones paying the price for poor MT quality. Translators and post-editors should therefore be seen as an integral part of the process, and the translator view taken into account in assessing MT quality.

Translation Quality Assessment forms a field of inquiry of its own within Translation Studies, but is mainly concerned with the assessment of professional translators (e.g. when applying for a post) and student translators during training (for an overview and benchmarking of frameworks, see O'Brien 2012). While the standard for human translation may be too high for machines, the common purposes for which machine translations are used – post-editing and gisting – set different quality criteria to those required for publication purposes. In these cases, a clumsy or even grammatically incorrect translation is good enough if the reader or translator can still interpret the meaning and edit the language as needed. The issue of differing quality criteria has been addressed, for example, in the form of guidelines by the Translation Automation User Society (TAUS 2010), and more recently as an International Standard draft (ISO/CD 18587:2014). Both these documents define two quality levels: 'good enough' quality and 'publishable quality' (TAUS 2010: 3-4), where 'good enough' is defined as comprehensible and accurate so that it conveys the meaning of the source text, but not necessarily grammatically or stylistically perfect. The draft standard (International Organization for Standardization 2014: 6) also specifies levels of post-editing ('light' and 'full') with goals similar to the two levels of the TAUS guidelines.

Most often the assumption for post-editing is a scenario where a bilingual post-editor corrects the MT based on the source text. However, some researchers have looked into a different scenario, monolingual post-editing. In the monolingual scenario, the post-editor either has no access to the source text or does not speak the source language. Real-world use cases for such situations might include crowdsourcing MT post-editing for emergency situations (Hu *et al.* 2011) or for online user forum posts – a case being studied by the European ACCEPT project. In such scenarios, the main question becomes whether the MT quality is good enough to convey the source text meaning and only linguistic corrections are necessary.

To explore this question, both Philipp Koehn (2010) and Chris Callison-Burch *et al.* (2010) adopted an approach where test subjects post-edited raw machine translations without access to the source text. The idea behind such an approach was to investigate whether current machine translation systems produce texts of sufficient quality for a monolingual reader to generate a translation without knowledge of the source text. The

correctness of the post-edited sentences was then evaluated based on a strict standard, with a correct sentence defined as "a fluent translation that contains the same meaning in the document context" (Koehn 2010: 541). While this standard is straightforward and corresponds to expectations for publication quality, it obscures the information about whether sentences are rated unacceptable for reasons of fluency or meaning. However, this distinction is important: not all errors are equally critical to the meaning of a sentence or a text as a whole, as already pointed out by Bensoussan and Rosenhouse (1990), and some prior studies have aimed to account for these different aspects. Hu *et al.* (2011), for example, investigated monolingual post-editing of text messages for emergency responders in a catastrophe situation. In this situation, an imperfect translation which conveys the meaning was acceptable. The translations were therefore evaluated on separate five-point scales for fluency of language and adequacy (preservation of meaning) with adequacy considered the more important criterion (Hu *et al.* 2011: 401-402). In their experiment comparing monolingual and bilingual post-editing of online user forum posts, Mitchell *et al.* (2013) evaluated quality with regard to fluency, comprehensibility and fidelity of meaning.

The distinction between fluency and meaning as well as differences between errors are also taken into account in some of the assessment systems in use in human translation, for example the systems used in Canada and Finland in certifying translators, which differentiate between translation and language errors and error severity rates (see Hale *et al.* 2012: 59-60 and Appendix 15, and Salmi and Penttilä 2013). On the MT side, a more detailed approach to error severity classification has been introduced by Irina Temnikova (2010). The classification is based on a previous MT error classification by Vilar *et al.* (2006), on studies in written language comprehension and error detection, and on Temnikova's earlier post-editing experiments. In this classification, ten error types are defined and ranked according to the presumed cognitive effort required to correct them, from 1 (easiest) to 10 (most difficult to correct). The easiest errors are considered to be connected to the morphological level, or correct words with incorrect form, followed by the lexical level, involving incorrect style synonyms, incorrect words, extra words, missing words and erroneously translated idiomatic expressions. The hardest errors in the classification relate to syntactic level and include wrong punctuation, missing punctuation, then word order at word level and finally word order at phrase level. Results reported in Temnikova (2010) suggest that pre-edited machine translations that had previously been found to require less post-editing effort measured by post-edit time and edit distance contain fewer errors that are cognitively more difficult when compared to MT that had not been pre-edited. A slightly modified version of this error difficulty scale was also used in Koponen *et al.* (2012) to investigate error types found in sentences with long or short post-editing times but with a similar number of errors. Sentences that took a long time to edit were found to contain more errors classified as cognitively difficult than those with short editing

times.

To further explore how well readers are able to interpret the source text meaning from a machine-translated text and perform in a monolingual post-editing situation, we adopted an approach similar to Koehn (2010) for the study described in this paper. Two short newspaper articles machine-translated from English into Finnish were post-edited by translator students acting as the test subjects without access to the source text, and both the machine translations and the edited versions were then assessed by the authors. Unlike in the previous studies, we chose to evaluate both correctness of meaning and correctness of language. The test subjects also rated the fluency and clarity of the text as well as its usability for post-editing or gisting.

2. Method

This section describes the study setup, the post-editing and evaluation task as well as our approach for evaluating the correctness of both the raw and edited versions.

2.1 Study setup

The material used for post-editing consisted of two English newspaper articles. The articles dealt with a study of the use of telecommunication technology (the 'telecom' text) and a new insect species discovered in Britain (the 'insect' text). They were written for the general public, requiring no previous knowledge of the subject matter. We chose these texts because of their general nature and also because they were texts used in Finnish high-school leaving examinations to measure foreign language reading comprehension, and could therefore be considered suitable for university students.

Both texts contained 673 words, with 32 sentences in the telecom text and 28 in the insect text. The texts were machine translated into Finnish using two MT systems: Google Translator, a statistical system based on large corpora of translated parallel texts and monolingual texts, and a system developed by Sunda Systems Oy, which is based on lexical rules, context rules, and syntactic rules. For editing, the machine translations were presented in a MS Word table with one sentence per row.

Text	**Source**	**Number of sentences**	**MT system**
telecom	Home truths about telecom *The Economist Technology Quarterly*, June 2007	32	Statistical
			Rule-based
insect	A scientific detective story *Time*, July 28, 2008	28	Statistical
			Rule-based

Table 1. The material used

The test subjects were translation students majoring in different languages (English, French, German, Italian, and Spanish) at the University of Turku who took part in an introductory course to translation technology. The course was taught in part by one of the authors. One of the topics discussed was machine translation (the different approaches, the MT systems available, its role in the translator's work). This task was one of the course assignments in which the students studied an example of the output of machine translation. Each test subject was assigned one of the four MT versions for editing. They were asked to correct the text into fluent and comprehensible Finnish according to their understanding of the meaning. The test subjects were also given the option of marking a sentence with 'nothing to correct' if they felt no corrections were needed, or 'unintelligible' if they felt unable to interpret the meaning of the sentence.

The students were not given the source text because we wished to measure the extent to which it is possible to obtain the correct meaning from a machine-translated text even though there are errors in the translation. This would not have been possible had they been given the source text. Also, this puts the students in the same position concerning the source text – as they were students of different languages, it could be assumed that the test subjects majoring in English (N=13) would have been more familiar with having an English source text than the others.

In addition, the test subjects were asked to rate on a five-point scale how fluent the text was and how clearly the meaning was conveyed. For an assessment of usability, the test subjects were asked whether the text was suitable for publication with no editing, post-editing without access to source text, post-editing with access to source text, gisting or unsuitable for any of these purposes. Some background information about the test subjects was also requested, such as their mother tongue and previous experience with machine-translated texts. At the end of the course, the students also submitted study journals reflecting on their course work, and journal comments regarding this task were collected for further information.

The number of test subjects editing each text and the number of edited sentences are shown in Table 2. Altogether, 48 students completed the assignment. Test subjects who indicated that Finnish was not their mother tongue were eliminated from this study. As only eight students editing the rule-based translation of the insect text completed the assignment (compared to 12 or 13 in the other groups), we requested additional

versions from students at the University of Helsinki, bringing the total for this group to 11.

Text	Source sentences	Test subjects	Sentences analysed	Edited sentences	Unedited sentences
Telecom-SMT	32	13	416	326	90
Telecom-RBMT	32	12	384	270	114
Insect-SMT	28	12	336	191	145
Insect-RBMT	28	11	308	249	59
Total	120	48	1,444	1,036	408

Table 2 Number of test subjects, edited sentences and unedited sentences

The 120 sentences edited by 48 test subjects produced 1,444 sentences for analysis. Of these, 72% (1,036 sentences) have been edited in some way by the test subjects. For the remaining 408 cases, one of the options 'nothing to correct' or 'unintelligible' had been used. There were also 11 cases where the test subjects had neither edited the sentence nor selected either of the two options.

2.2 Evaluation of correctness

We evaluated both the raw machine translations and the edited sentences for correctness of meaning and language. First, different approaches were considered for the correctness evaluation. A scalar evaluation of different aspects, as described in Fiederer and O'Brien (2009) or Hu *et al.* (2011), was considered, but with a relatively high number of sentences to be evaluated (120 raw MT sentences and 1,444 post-edited sentences) a more straightforward binary correct/incorrect scale like that of Callison-Burch *et al.* (2010) and Koehn (2010) appeared more feasible. In addition, we wanted to differentiate incorrect use of target language from actual translation errors that affect meaning. One of the authors was already familiar with the system used in evaluating the translation assignments for the system of certifying translators in Finland (Salmi and Penttilä 2013), and had used it in her translation courses at the University of Turku. Therefore, we decided to evaluate correctness of meaning and correctness of language on separate binary scales to produce four categories:

- correct meaning – correct language: the sentence correctly conveys the meaning of the source text and is grammatically and idiomatically correct
- correct meaning – incorrect language: the sentence correctly conveys the meaning of the source text despite errors in spelling, punctuation, grammar or word choice
- incorrect meaning – correct language: the sentence fails to convey source text meaning or conveys a different meaning but is

grammatically and idiomatically correct
- incorrect meaning – incorrect language: the sentence fails to convey the source text meaning and contains errors in spelling, punctuation, grammar or word choice

The evaluation was carried out by the two authors, who are both native Finnish speakers with experience in professional translation from English and who have both experience in translator training. The evaluation was first made independently, and then combined. The first assessments matched in about 65% to 70% of the sentences. Of the differing cases, about 10% were explained by minor typographical errors escaping the notice of one of the evaluators. Cases where the two evaluators differed over whether a certain word choice or grammar issue should be considered correct (even if clumsy), incorrect language or even incorrect meaning, were discussed and a final evaluation was agreed upon for each case. Assessments given for different versions of the same sentence were also checked for consistency.

3. Results

Section 3.1 presents the results of the correctness evaluation of the raw machine translations and the sentences edited by the test subjects. Section 3.2 discusses the types of errors found and their effect on the success of corrections. The results of the fluency, clarity and usability rating by the students have been discussed in more detail in Koponen and Salmi (2012).

3.1 Correctness analysis

Figure 1 shows the overall correctness evaluation of the raw machine translations and the post-edited versions. In the raw MT state (left column of Figure 1), only six of the 120 sentences (5%) were evaluated as fully correct, and a further 24 sentences (20%) were evaluated as correct with regard to meaning but with language errors. Some differences appeared between the two systems, in that the rule-based system produced 21 of the 30 sentences judged to convey the correct meaning, whereas the statistical system produced only nine. Nearly all sentences contained language errors, with only three sentences (2.5%) assessed as incorrect for meaning but grammatically correct, and the majority of cases (87 sentences or 72.5%) were assessed as incorrect with regard to both language and meaning. Between the two systems, the rule-based system produced slightly more cases that were free of language errors.

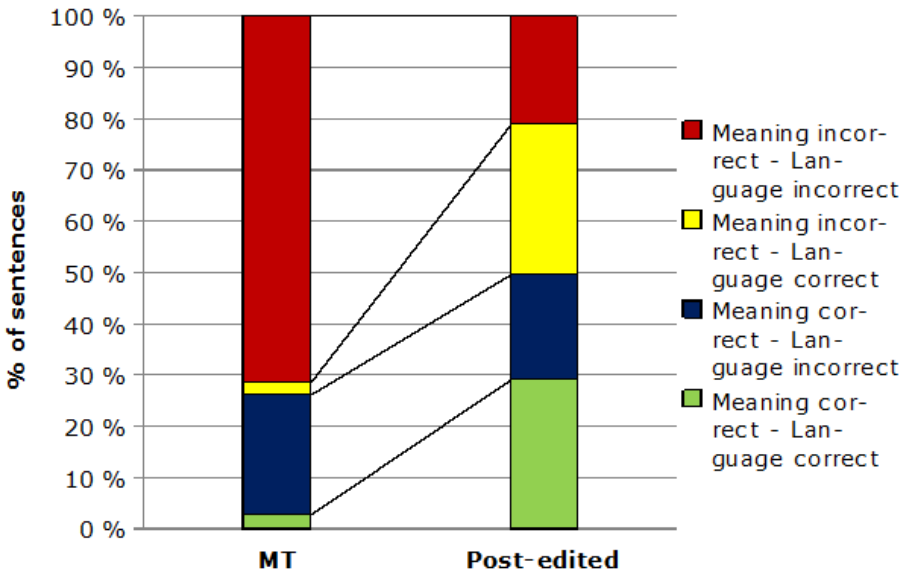

Figure 1. Evaluation of correctness in raw MT and post-edited versions

After post-editing (right column of Figure 1), 306 cases (29.5% of the 1036 post-edited sentences) are judged fully correct. These cases include only sentences successfully edited by the test subjects. Sentences that were already assessed as fully correct in the MT and recognised as such by the test subjects using the option 'nothing to correct' are discussed below. A considerable number of sentences still contain language errors after post-editing. A further 210 cases (20.3% of the post-edited sentences) were considered to contain the correct meaning but with language errors. Often, these are grammatical errors or non-idiomatic word choices in the MT left unedited, but there are many cases where the test subjects themselves have introduced new errors that range from typographical to clear grammatical errors. However, as our main focus was on the correctness of meaning, the success of post-editing is determined here as the number of sentences with correct meaning after editing. All sentences with correct meaning can therefore be considered successful although some language errors remain – overall, these account for 49.8% of the cases.

On the other hand, all edited sentences with incorrect meaning are unsuccessful regardless of the correctness of language. In addition to edited sentences, unedited sentences also factor in the success rate. The number of cases where the options 'nothing to correct' and 'unintelligible' have been used are shown in Figure 2. All sentences labelled 'unintelligible' are unsuccessful. Although nearly all of these are sentences with both incorrect meaning and incorrect language (302 of the total 331 uses of 'unintelligible'), in some cases a sentence with correct meaning has been deemed unintelligible (three cases of fully correct sentences and 18 cases of sentences with correct meaning but incorrect language). For 'nothing to correct,' success depends on the raw MT assessment: these cases are only successful if the MT sentence has the correct meaning. Most often, this

option is used correctly (32 cases of fully correct, 25 with language errors), but there are nine cases where the MT meaning is actually incorrect (one with correct language, eight with language errors).

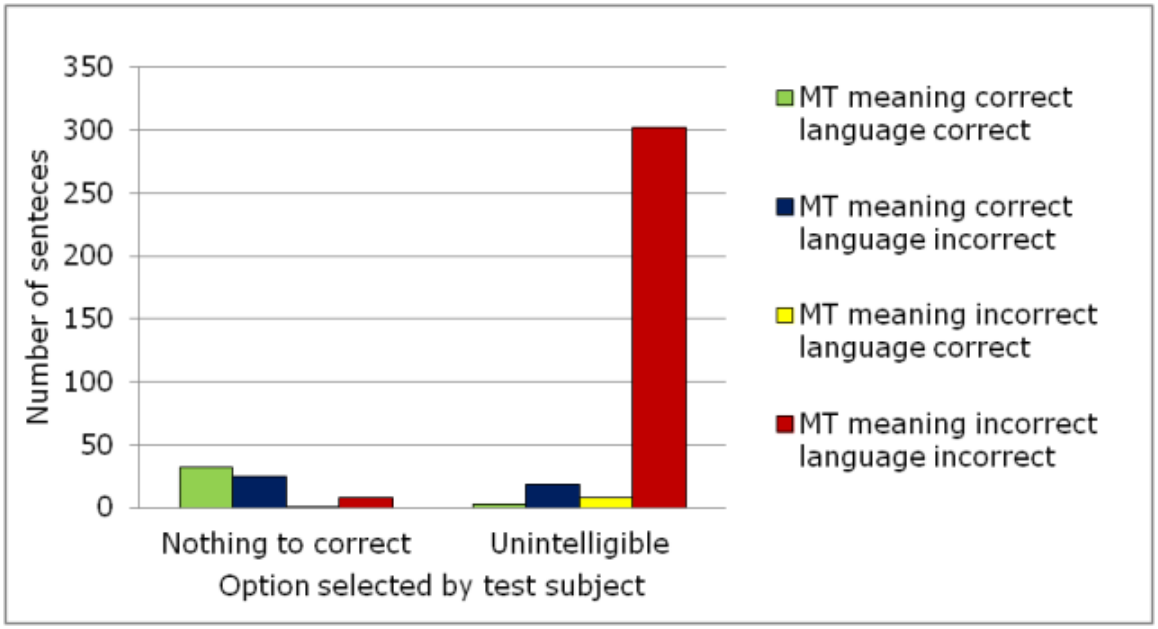

Figure 2. Unedited sentences

We can also examine post-editing success compared to the original raw MT assessments. Of the cases judged to have correct meaning after post-editing (with or without language errors), some involved cases where the meaning was already correct in the MT (22 fully correct in MT and 191 with only language errors), but the majority involve cases where the MT sentence had been judged incorrect for meaning (eight with correct language, 295 with both incorrect meaning and incorrect language). On the other hand, nearly all of the cases where the sentence was judged incorrect for meaning after post-editing are cases where the meaning of the MT sentence was also incorrect (17 with correct language, 443 with incorrect language). However, there are a few cases where the test subject had edited a sentence with originally correct meaning in some way that changed it to incorrect (eight fully correct, 52 with language errors).

Overall, many sentences have mixed success in that some test subjects are able to edit them correctly while others are not, but some sentences can be identified as particularly easy or particularly difficult. Figure 3 shows the total numbers of sentences divided into four 'success categories.' The 31 sentences that have been successfully edited (or recognised as correct) by none or only one test subject (< 10%) in the group can be considered most difficult. In contrast, the 15 sentences that have been successfully edited by all test subjects, or all but one, (> 90%) are the easiest sentences. To investigate cases where editing is easy or difficult, a more detailed analysis of these two sets of sentences is presented in the next section.

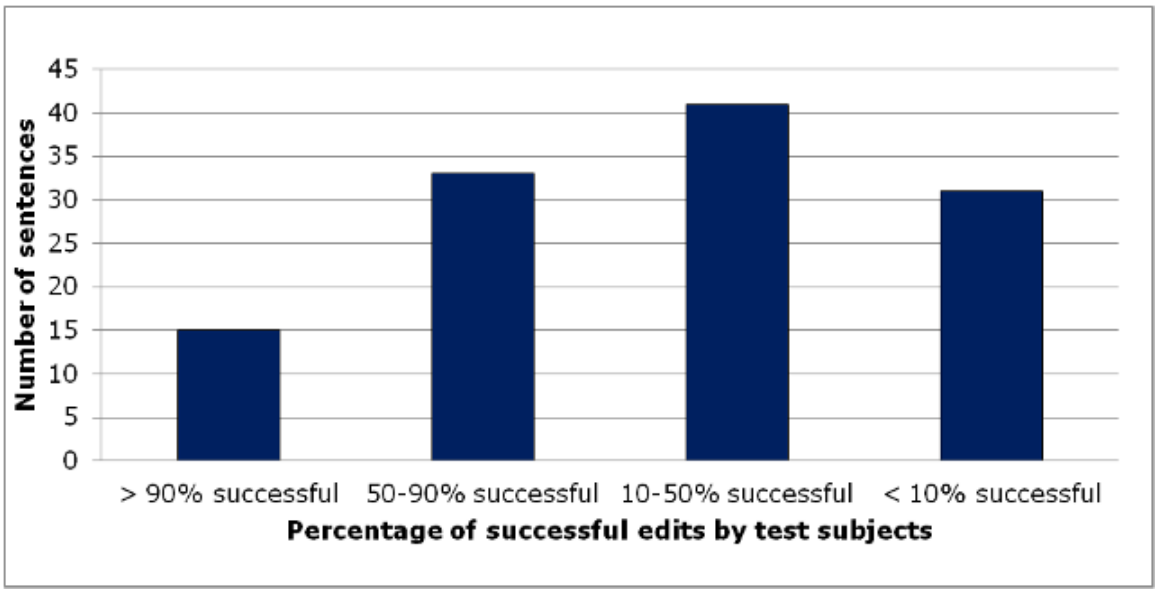

Figure 3. Number of sentences with different success rates

3.2 Error analysis of easy and difficult cases

To explore the difference in ease of correction, we performed a more detailed error analysis on the 15 sentences identified as particularly successful and the 31 sentences identified as particularly difficult. Of the 15 successful sentences, four were correctly translated by the machine translation system, and they were left out of the analysis. To analyse, we used Temnikova's (2010) classification into error types on morphological, lexical and syntactic levels. In addition to the error types she used, we found it useful to add one error type on the morphological level, namely typographical errors (for example, capitalization errors or case endings that break the Finnish vowel harmony but are recognised as the correct case), one error type on the lexical level, namely words left in the text in source language (untranslated), and another error type that shows itself on a lexical level but is, in fact, related to the syntax: incorrect word where the machine translation has changed the part of speech of the word (for example, a verb has been translated as a noun, or an adjective as an adverb). We also divided incorrect words into two types: incorrect words in correct form and incorrect words in incorrect form.

An overall quantitative analysis of the 42 sentences shows that the particularly unsuccessful sentences contain more words (24 on average) and more errors (7.6 on average) than the particularly successful sentences (13 words and 2.6 errors on average). The errors in the successful sentences are mainly on the morphological level (incorrect word forms and typographical errors) and on the lexical level (incorrect style synonyms), and the successful sentences contain no errors related to idioms or word order. The errors in the unsuccessful sentences are mainly related to lexical level (incorrect and missing words), and they also contain errors related to word order and idioms.

A simple example of a successfully corrected sentence is the following:

Example 1
ST: The spider turned out to be quite *common* – in the Canary Islands.
MT: Hämähäkki osoittautui melko *yleisiä* – Kanariansaarilla.
'The spider turned out to be quite *common (plural indefinite form)* – in the Canary Islands.'
(Insect-SMT, 11/12 successful, 1/12 unsuccessful)

In Example 1, only the adjective has been translated in an incorrect, plural form and is easy to correct. On the other hand, even multiple errors in the same sentence can sometimes be successfully corrected as in the case of Example 2, where five errors were identified:

Example 2
ST: Last year at a *food-processing* factory near Geneva, the workers revolted when the director tried to ban *mobile phones from* the factory floor, and *he was forced* to relent.
MT: Viime vuonna *food-processing?*-tehtaassa Geneven lähellä työntekijät kapinoivat, kun johtaja yritti kieltää *matkapuhelimia pääsemästä* tehdassaliin, ja *hän oli pakotettu* antamaan periksi.
'Last year at a *food-processing?* factory near Geneva the workers revolted when the director tried to ban *mobile phones from accessing* the factory floor, and *he was compelled* to relent.'
(Telecom-RBMT, 12/12 successful, 0/12 unsuccessful)

The machine translation system has not been able to translate 'food-processing' and has left it in English; the system marks all untranslated words with a question mark. These errors are categorised as untranslated word and incorrect punctuation. The translation of 'he was forced to relent' is too formal for a newspaper text register, and has been marked as a stylistic error. In addition, the banning of mobile phones has been translated in a way that gives the mobile phones an active role. These errors are categorised as an incorrect word and an incorrect word order (*matkapuhelimia pääsemästä* should have been translated as *tuomasta matkapuhelimia*, lexical level). However, they are easy to correct based on the context and general world knowledge.

In one case, the two different versions of the same sentence appear on the list of both successfully and unsuccessfully corrected sentences, which provides an interesting comparison. The easy case, Example 3a below, contains three errors: one typographical error (*Barclaytä* should be *Barclayta*), one incorrect form (*latinalainen* 'Latin') which makes its relation to the rest of the sentence unclear, and one mistranslated word: 'proper name' (in this context, 'appropriate' or 'correct name') has become *erisnimi* 'proper noun.' In spite of this error on lexical level, the test subjects have probably been helped by their knowledge of the fact that species are generally given Latin names, and all but one have edited the sentence

correctly.

Example 3
ST: But that won't stop scientists like Barclay from trying to give his new chums a *proper name* – that is to say, *a Latin one*.
Translation 3a:
MT: Mutta se ei estä tiedemiehiä kuten Barclaytä yrittämästä antaa hänen uusille kavereillensa *erisnimen* – toisin sanoen, *latinalainen*.
'But that won't stop scientists like Barclay from trying to give his new chums a *proper noun* – that is to say, *Latin*.'
(Insect-RMBT – 10/11 successful, 1/11 unsuccessful)

Quite interestingly, however, the same sentence translated by the statistical system ranks among the particularly unsuccessful ones:

Translation 3b:
MT: Mutta se ei lopeta tiedemiehet Barclay yrittävät antaa hänen uusi ystävämme oikea nimi - toisin sanoen, Latinalaisen yksi.
'But it does not end scientists Barclay try (third person plural) to give his new our friend proper name – in other words, Latin's one.'
(Insect-SMT – 1/12 successful, 7/12 unsuccessful, 4/12 not edited)

This translation contains 11 errors. There are three errors on the lexical level (two incorrect words – *lopeta* for 'stop' and *yksi* for 'one,' and one missing word – 'like'), and eight on morphological level: seven words in incorrect forms (*tiedemiehet* for *tiedemiehiä*, *yrittävät* for *yrittämästä*, *uusi* for *uudelle*, *ystävämme* for *ystävällemme*, *oikea* for *oikeaa*, *nimi* for *nimeä*, *Latinalaisen* for *latinalainen*) and one typographical error (*Latinalainen* spelled with a capital). Although the sentence and even most of the error types are the same as in 3a, this second version contains more errors, leading to a completely different effect. It is probably the incorrect verb form of 'from trying,' translated as 'they try,' together with the missing translation of 'like' that make it too difficult for most of the subjects to figure out how the rest of the words relate to each other, even though they appear in the right order.

Analysis of the particularly difficult sentences showed that some of the errors considered cognitively more difficult in Temnikova's (2010) classification have, in fact, been impossible for the test subjects to recover from. These include missing words and idiomatic expressions. Example 4 shows a sentence where none of the test subjects has been able to figure out the correct meaning:

Example 4
ST: Her research in Switzerland and France found that even when *people are given unlimited cheap or free calls*, *the number and length of calls does not increase* significantly.
MT: Hänen tutkimuksensa Sveitsissä ja Ranskassa havaitsi, että jopa

silloin kun *ihmisille soitetaan*, *ei kasva* merkittävästi.
'Her research in Switzerland and France found that even when *people are called*, *does not increase* significantly.'
(Telecom-RMBT – 7/12 unintelligible, 0/12 successful, 5/12 unsuccessful)

In total, ten words are missing from the translation and three words have been mistranslated. The omissions of both 'unlimited cheap' or 'free and *the number and length of calls'* and the mistranslation of 'people are given [...] calls' to *ihmisille soitetaan* 'people are called' change the meaning significantly. While the omission in 'does not increase' is clear, no clues remain here of the correct meaning. Furthermore, there is nothing to indicate that the other part ('people are called' instead of 'people are given free and unlimited calls') is actually incorrect. On the contrary, it appears to fit well into the sentence and none of the test subjects changed this part at all.

In some cases, even one error in an idiomatic expression may severely affect the sentence. One such case appears in the insect text, where the rule-based system has 'understood' the idiom '(noun) and a half' to mean 'quite a (noun)' rather than literally 'decade and a half' (15 years) and translated this as *aikamoinen vuosikymmen* 'quite a decade.' In this case, nine out of 11 test subjects have edited the sentence (one has chosen 'unintelligible' and one 'nothing to correct'), for example by deleting *aikamoinen* ('quite a') but no one has been able to guess that the length of time is wrong. Another case, where a missing word and an error in one word form significantly change the entire sentence, is given in Example 5.

Example 5
ST: Barclay is not convinced that climate change is *responsible for Britain's new inhabitants*.
MT: Barclay ei ole vakuuttunut siitä, että ilmastonmuutos on *vastuussa Britannian asukkaille*.
'Barclay is not convinced that climate change is *accountable to Britain's inhabitants*.'
(Insect-SMT – 6/12 'unintelligible,' 1/12 successful, 5/12 unsuccessful)

The wrong word form *asukkaille* 'to the inhabitants' changes the meaning so that while it is grammatical, the sentence makes no sense in this context (possibly any context), and this appears to have been recognised by the test subjects. Some have omitted 'Britain's inhabitants' completely while others have tried to add something, for example 'responsible for the concerns of Britain's inhabitants', and there is even one edit where the meaning changes completely to 'Barclay is not convinced that Britain's inhabitants are responsible for the climate change.'

Not all our findings systematically follow the error ranking suggested by Temnikova (2010). In our data, errors in punctuation appear in both successful and unsuccessful sentences, and they alone do not seem to make

the MT output particularly difficult to understand. Many of the punctuation errors in our data appear fairly simple, such as extra quotation marks, and cases where punctuation errors are the only ones in a sentence are generally successfully corrected. In contrast, the seemingly easy errors in word forms can sometimes be crucial. A particular case of incorrect words or incorrect word forms are words where the grammatical function (or part of speech) has changed in the translation. One such case of incorrect form changing the meaning drastically appears in Example 5 above. In another one of the particularly unsuccessful sentences of the telecom text, the rule-based system had rendered the verb in the phrase 'She based her research on […] asking people to keep logbooks,' where 'asking people' is a construction of a verb and a noun, as *kysyminen kansoittaa* ('an/the asking populates'), a construction with a noun and a verb in the third person singular form. The sentence contained ten errors altogether, and none of the test subjects were able to guess the correct meaning.

4. Discussion

Overall, the raw machine translations rated quite poorly for correctness, as only a few sentences were judged fully correct or even correct with regard to meaning but not with regard to language. For post-edited sentences, 29.5% were judged fully correct with regard to both meaning and language. This result is comparable to prior studies, where results for different systems and language pairs have ranged from 26% to 35% (Koehn 2010) and from 10% to 80% (Callison-Burch *et al.* 2010). When cases with correct meaning but incorrect language are considered, editing the meaning was successful in just under 50% of the cases. This can be compared to Hu *et al.* (2011) who also assessed meaning separately from language, and defined 'fully correct' as two evaluators giving the highest adequacy score (5) for a sentence. In their results, 24% to 39% of sentences (depending on system and test set) were rated as fully correct. In the study reported by Mitchell *et al.* (2013), the editors were able to achieve similar levels of fluency and comprehensibility with and without the source text, but editing monolingually without the source text led to less fidelity, meaning that the editors were not able to fully recover the entire meaning of the source text.

The results of our study also show that the test subjects have been quite inattentive to language errors. Just as discovered by Koehn (2010), there are quite a few cases where only some of several errors within a sentence have been edited or the errors have been ignored and 'nothing to correct' has been used. Common examples are minor punctuation errors or the English title *Ms* remaining in the Finnish translation. Furthermore, sometimes the test subjects have made language errors that were not present in the machine translation. Most seem to be simple typographical errors, but sometimes grammatical errors have also been introduced, for example by editing a part of the sentence and failing to correct the rest accordingly. In their study journals, many students noted that the number of errors made editing cumbersome and that they would not be very happy

with such work, which may have played a role.

The analyses of particularly successfully and particularly unsuccessfully edited sentences in Section 3.2 suggest some differences between the kind of errors that can be easily recovered from and errors that are more critical to understand the meaning. Overall, long sentences and sentences with multiple errors appear difficult, which has also been observed in other studies (e.g. Tatsumi 2009; Koponen 2012). In certain cases, very short segments, such as titles, can also be difficult, as observed by Tatsumi (2009). In line with Temnikova's (2010) suggestions, errors in the word form are the easiest errors to recover from, and errors in missing words, idiomatic expressions and word order the hardest. Examples of relatively easy cases are the ones where certain correct words have been rendered in incorrect form but the relations between words in the sentence can be deduced based on context and general knowledge (Examples 1 and 3a). Errors that appear on a lexical level and omitted idioms (Example 5) render the test subjects either unable to correct the error successfully or unable to edit the error at all. However, the ranking of error gravity does not always appear to be as straightforward as Temnikova (2010) suggests, as illustrated by the relative easiness of punctuation errors on the one hand, and the crucial role word forms may sometimes play on the other. Even the errors assumed to be the easiest – correct words in incorrect forms – may turn out to be difficult to correct when they severely impact the syntactic relations of the sentence. Similar observations were made by Koponen *et al.* (2012) with regard to error types found in sentences with long vs. short editing times. However, a more thorough analysis of all our data would be necessary in order to suggest changes in Temnikova’s error classification.

Editing fatigue could have affected the test subjects' willingness to edit or the success of editing, but no correlation was found between the position of the sentence in the text and the number of test subjects editing successfully or unsuccessfully. Some further steps toward assessing the workload of post-editing perceived by translation students were taken in another study where we asked translation students to translate into Finnish the first paragraph of the insect text using two different MT systems and evaluate the task. A third of the 49 students answered they would have translated the text faster themselves, another third felt there would not have been a difference in time, and a fifth considered MT with post-editing faster (nine did not comment). Half (25) of the students also commented that post-editing was easy, did not take long or did not demand much effort; nine found it difficult and 15 did not comment (Salmi and Koponen 2014).

Prior studies (Callison-Burch *et al.* 2010; Koehn 2010) have also discovered considerable variation between individual test subjects, which was also noticeable in this study. Some test subjects were particularly successful, and the overall best result was 26 sentences out of 32 correct (Telecom-SMT). On the other hand, some test subjects achieved much less success than others, the lowest number of correct sentences being four out of 28

(Insect-SMT). The differences between test subjects have been examined more closely in Koponen and Salmi (2012).

5. Conclusion

This paper presented the results of a machine translation post-editing task where test subjects attempted to edit a raw machine translation without access to the source text, to measure the extent to which it is possible to obtain the correct meaning from a machine-translated text even though there are errors in the translation. The results show that the correctness of raw machine translations was not rated highly (30 out of 120 machine translated sentences deemed to convey the correct meaning with or without language errors), and in post-editing the text, the test subjects were able to arrive at the correct meaning based on the machine translation alone in approximately half of the sentences that had been post-edited.

In the correctness assessment procedure, differentiating between correctness of meaning and correctness of language proved useful. Since the test subjects were not particularly attentive to language errors and even introduced new errors, a simple binary scale (correct/incorrect) without separation of language and meaning would not have reflected their ability to deduce the meaning. Nonetheless, this assessment remains relatively coarse, and combining it with a more detailed error analysis of particularly easy and particularly difficult cases provided more information.

The results indicate that long sentences and sentences with a high number of errors compared to the sentence length are more difficult to edit. The detailed analysis of particularly difficult and particularly easy cases also suggests some error types that are easy or difficult to recover from. The easy errors include changes in the word form when the correct form can be deduced based on the context in the sentence. The difficult errors are especially missing words, incorrectly translated words and idiomatic expressions. In some cases, even one single error can change or obscure the meaning of a sentence so severely that its meaning cannot be recovered.

Further work will concentrate on finding more accurate ways of defining and identifying errors that are particularly critical to the meaning of the sentence or the whole text. We will also attempt to further develop the assessment procedure, for example by more clearly defining the lines between language errors and translation errors and using language technology tools to support the manual evaluation. The students' performance and subjective assessments of the fluency, clarity and usability of the texts will also be explored in greater detail.

Bibliography

- **Bensoussan, Marsha and Judith Rosenhouse** (1990). "Evaluating student translations by discourse analysis." *Babel,* 36(2), 65-84.

- **Callison-Burch, Chris *et al.*** (2010). "Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation." *ACL 2010: Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. Proceedings of the workshop,* 17-53. http://www.statmt.org/wmt10/pdf/wmt10-overview.pdf (consulted 12.09.2014)

- **Carl, Michael *et al.*** (2011). "The process of post-editing: A pilot study." *Proceedings of the 8th international NLPSC workshop. Special theme: Human-machine interaction in translation* (Copenhagen), 131-142.

- **Fiederer, Rebecca and Sharon O'Brien** (2009). "Quality and machine translation: A realistic objective?" *The Journal of Specialised Translation* 11, 52-74. http://www.jostrans.org/issue11/art_fiederer_obrien.pdf (consulted 12.09.2014)

- **Hale, Sandra *et al.*** (2012). *Improvements to NAATI testing. Development of a conceptual overview for a new model for NAATI standards, testing and assessment. Report for The National Accreditation Authority for Translators and Interpreters (NAATI)*. http://www.naati.com.au/PDF/INT/INTFinalReport.pdf (consulted 12.09.2014)

- **Hu, Chang *et al.*** (2011). "The value of monolingual crowdsourcing in a real-world translation scenario: simulation using Haitian Creole emergency SMS messages." *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT '11)*, 399-404. http://www.statmt.org/wmt11/pdf/WMT48.pdf (consulted 12.09.2014)

- **International Organization for Standardization** (2014). *ISO/CD 18587:2014 Translation Services – Post-editing of machine translation output – Requirements*.

- **Koehn, Philipp** (2010). "Enabling monolingual translators: Post-editing vs. options." *NAACL HLT 2010: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. *Proceedings* (Los Angeles, California, June 2010), 537-545. http://aclweb.org/anthology-new/N/N10/N10-1078.pdf (consulted 12.09.2014)

- **Koponen, Maarit** (2012). "Comparing human perceptions of post-editing effort with post-editing operations." *Proceedings of the Seventh Workshop on Statistical Machine Translation* (Montréal, Canada, June 2012), 181–190. http://www.aclweb.org/anthology-new/W/W12/W12-3123.pdf (consulted 12.09.2014)

- **Koponen, Maarit *et al.*** (2012). "Post-editing time as a measure of cognitive effort." *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice* (San Diego, USA, October 2012). http://amta2012.amtaweb.org/AMTA2012Files/html/13/13_paper.pdf (consulted 12.09.2014)

- **Koponen, Maarit and Leena Salmi** (2012). "Does prior use of machine translation systems help in post-editing?" Paper presented at *International Workshop on Expertise in Translation and Post-editing: Research and Application* (Copenhagen Business School, 17-18 August 2012).

- **Mitchell, Linda, Johann Roturier and Sharon O'Brien** (2013). "Community-based post-editing of machine-translated content: monolingual vs. bilingual." *Workshop Proceedings: Workshop on Post-editing Technology and Practice (WPTP-2)*, 35-44.

- **O'Brien, Sharon** (2012). "Towards a dynamic quality evaluation model for translation." *The Journal of Specialised Translation* 17, 55-77. http://www.jostrans.org/issue17/art_obrien.pdf (consulted 12.09.2014).

- **Plitt, Mirko and François Masselot** (2010). "A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context." *The Prague Bulletin of Mathematical Linguistics* 93, 7-16. http://ufal.mff.cuni.cz/pbml/93/art-plitt-masselot.pdf (consulted 12.09.2014)

- **Robert, Anne-Marie** (2013). "Vous avez dit post-éditrice ? Quelques éléments d'un parcours personnel." *The Journal of Specialised Translation* 19, 29-40. http://www.jostrans.org/issue19/art_robert.pdf (consulted 12.09.2014)

- **Salmi, Leena and Maarit Koponen** (2014). "Machine Translation, Post-Editing and Respeaking: Challenges for Translator Training." *Man vs. Machine? The Future of Translators, Interpreters and Terminologists. Proceedings of the XXth FIT World Congress Vol. I*, (Berlin, August 2014), 138-145.

- **Salmi, Leena and Ari Penttilä** (2013). "The System of Authorizing Translators in Finland." Dina Tsagari and Roelof van Deemter (eds). *Assessment Issues in Language Translation and Interpreting*. Frankfurt: Peter Lang, 115-130.

- **Tatsumi, Midori** (2009). "Correlation between Automatic Evaluation Metric Scores, Post-Editing Speed, and Some Other Factors." *MT Summit XII: Proceedings of the twelfth Machine Translation Summit*, (Ottawa, Canada, August 2009), 332-339. http://www.mt-archive.info/MTS-2009-Tatsumi.pdf (consulted 27.08.2013).

- **TAUS** (2010). *Machine Translation Post-editing Guidelines*. https://evaluation.taus.net/resources/guidelines/post-editing/machine-translation-post-editing-guidelines (consulted 09.05.2014).

- **Temnikova, Irina** (2010). "Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment." *LREC 2010: Proceedings of the seventh international conference on Language Resources and Evaluation*, (Valletta, Malta, May 2010), 3485-3490. http://www.lrec-conf.org/proceedings/lrec2010/pdf/437_Paper.pdf (consulted 12.09.2014).

- **Thicke, Lori** (2013). "The industrial process for quality machine translation." *The Journal of Specialised Translation* 19, 8-18. http://www.jostrans.org/issue19/art_thicke.pdf (consulted 27.08.2013).

- **Vilar, David *et al.*** (2006). "Error analysis of statistical machine translation output." *LREC-2006: Fifth International Conference on Language Resources and Evaluation. Proceedings*, (Genoa, Italy, May 2006), 697-702. http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf (consulted 27.08.2013).

Websites

- http://www.accept-project.eu/

- http://translate.google.com

- http://www.sunda.fi/eng/products.html

Biographies

Maarit Koponen has an MA in English Philology (2005) from the University of Helsinki and is currently writing her PhD in Language Technology (focusing on machine translation evaluation and post-editing) at the University of Helsinki, Department of Modern Languages. She works as a Translation Studies teaching assistant and has taught computer-assisted translation at the University of Helsinki and professional translation as a visiting teacher at the University of Eastern Finland. She has also worked as a professional translator for over five years.

E-mail: maarit.koponen@helsinki.fi

Dr Leena Salmi currently works as Professor of Multilingual Translation Studies at the School of Languages and Translation Studies at the University of Turku. She also works as a freelance translator. She has lectured in topics such as translation technology, localisation and EU translation. Her PhD (2004) dealt with the usability of computer documentation, and her current research interests include information-seeking on the Web as part of the translator's work and the amount of translated texts in everyday life.

E-mail: leena.salmi@utu.fi