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

Joke Daems **Lieve Macken** **Sonia Vandepitte**
Department of Translation, Interpreting and Communication
Ghent University
Belgium
firstname.lastname@ugent.be

Abstract

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

1 Introduction

With the translation industry exponentially growing, more hope is vested in the use of machine translation (MT) to increase translators' productivity (Rinsche and Portera-Zanotti, 2009). Though post-editing MT has proven to increase productivity and even quality for certain text types (Tatsumi, 2010), research on the usability of post-editing for more general texts is rather limited. The research presented in this paper is a pilot study conducted as part of the ROBOT-project[1], a project designed to gain insight in the differences between human translation and the post-editing of machine translation. The process and product of translation are the two main areas of interest of the project, and results of student translators and professional translators shall be compared. In this paper, the translation quality assessment approach developed for the project is presented and tested on translations of student translators. This fine-grained, two-step approach does not only allow for the analysis and comparison of translation problems for different methods of translation (such as human translation and post-editing), but can also be used as an evaluation method for different text types and goals. As such, it is a useful tool, both for researchers and people concerned with the evaluation of translation quality in general.

2 Related Research

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

[1] lt3.hogent.be/en/projets/robot

Sharon O'Brien, Michel Simard and Lucia Specia (eds.)
Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice, Nice, September 2, 2013, p. 63–71.

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

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

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

More recently, researchers have tried to overcome the flaws of previous metrics by paying more attention to the goal of the source text and target text. Williams (2009) suggests using an argumentation centred translation quality assessment to make sure the macro structure of the source text is respected, whereas O'Brien (2012) opts for a dynamic approach to quality evaluation, taking into account the goal of the translation, the time and the resources of the company. Colina (2009) introduces an assessment tool with a functionalist approach, allowing for a user-defined notion of quality. Though the idea behind the tool is in line with the need for a more flexible approach to translation quality assessment, the tool's usefulness is limited to providing a quick assessment of the main problem categories of a text, but it lacks an in-depth analysis of the types of errors and gives no concrete suggestions for improvement.

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

3 A two-step TQA approach

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

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

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

3.1 Categorization

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

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

Table 1: Adequacy subcategories with error scores

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

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

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

3.2 Objectivity & Flexibility

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

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

Table 2: Acceptability subcategories with error scores

user-defined categorizations and error weights.

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

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

4 Experiment

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

4.1 Experimental set-up

Participants were 16 Master's students of translation taking a general translation course. Students had no experience with post-editing and were given no specific training. Each student received a translation and a post-editing task. The machine translation to be post-edited was obtained by using Google Translate[2]. The order in which the students received the tasks differed, to reduce task order effects. There was no time restriction. The corpus consisted of four newspaper articles of more or less equal length (260-288 words) taken from the Dutch Parallel Corpus (Macken et al., 2011). The instruction for both tasks was to achieve a translation of publishable quality, and the target audience of the translations was said to be more or less equal to the target audience of the source text. Participants were informed that they would receive feedback on the productivity and quality of their translations.

[2]translate.google.com

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

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

4.2 Translation Quality Assessment

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

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

4.3 Inter-annotator agreement

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

A consolidation phase was introduced to check

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

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

4.4 Results

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

4.4.1 Productivity

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

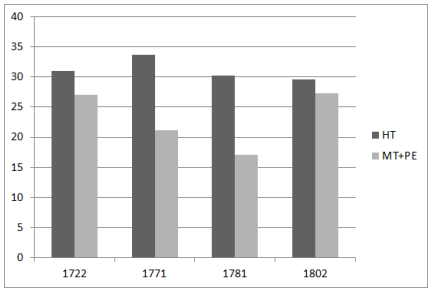

Figure 1: Time spent per text in minutes

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

4.4.2 Quality: totals

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

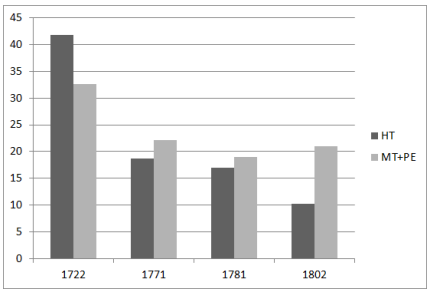

Figure 2: Average acceptability error score per text

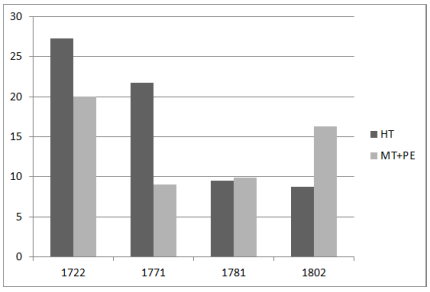

Figure 3: Average adequacy error score per text

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

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

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

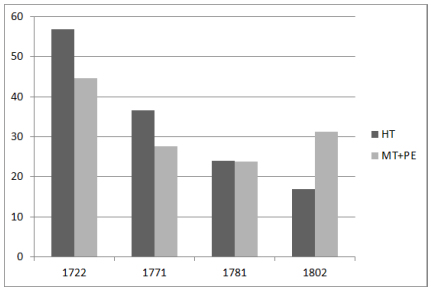

Figure 4: Average total error score per text

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

4.4.3 Quality: problem analysis

Though the totals for adequacy and acceptability already provided some insights in the differences between post-editing and human translation, the main goal of the proposed categorisation was to allow for a more thorough analysis of translation problems. A more detailed picture is given by the analysis of the main subcategories. As was the case for the difference between adequacy and acceptability, the scores for each category depend largely on the text. When looking at the most common problems for each text, it becomes clear that 'meaning shift - other' and 'meaning shift - deletion' are very common categories for human translation for texts 1722 and 1771 (with 'other' taking up 9% and 15% of total human translation error for these texts and 'deletion' accounting for 7% and 15% of total human translation errors), but not so common for post-editing (the categories 'other' and 'deletion' were not found in text 1722 and only accounted for 3% and 5% of all post-editing errors for text 1771, respectively). Common post-editing problems, on the other hand, seem to be 'meaning shift - wrong word sense' and 'lexicon - wrong collocation'. 'Wrong word sense' accounted for 14% of all PE-errors for text 1722 (for HT, this was a mere 5%), 9% of all PE-errors for text 1781 (versus 1% for HT), and 10% of all PE-errors for text 1802 (compared to 4% for HT). 'Wrong collocation' accounted for 17% of all PE-errors for text 1722 (compared to 7% for HT), 9% of all PE-errors for text 1771 (compared to 5% for HT) and 17% for text 1781 (compared to 7% for HT). Some categories were only important issues for one of the four texts, such as 'compounds' for text 1722 (with a HT and PE value of 4% and 8% respectively), 'capitalization' for text 1771 (with a HT and PE value of 8% and 4% respectively) and 'register' for text 1802 (with a HT and PE value of 7% and 1% respectively). A more thorough analysis of these differences could yield insights in text differences and what makes a text difficult to translate (both for a human and a machine), but this would go far beyond the scope of the present article.

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

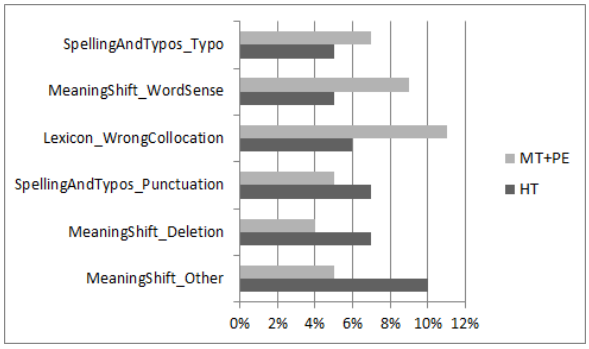

Figure 5: Most common error types over all texts

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

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

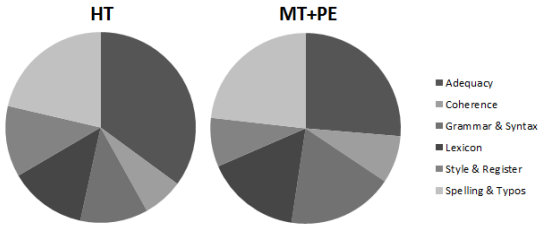

Figure 6: Proportion of problem categories: focus on acceptability

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

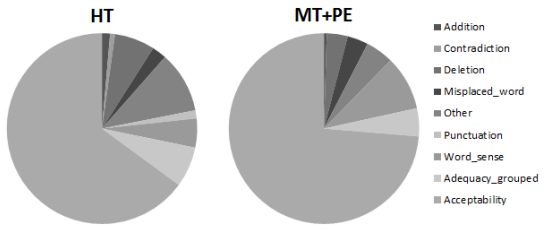

Figure 7: Proportion of problem categories: focus on adequacy

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

5 Discussion & Future work

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

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

References

Aziz, Wilker, Sheila de Sousa, and Lucia Specia. 2012. PET: a tool for post-editing and assessing machine translation. *LREC 2012, The 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey. May 2012.

Chesterman, Andrew. 1998. *Causes, Translations, Effects*, Target, 10(2):201-230.

Colina, Sonia. 2009. *Further Evidence for a Functionalist Approach to Translation Quality Evaluation*, Target 21(2):235-264.

Daems, Joke, and Lieve Macken. 2013. *Annotation Guidelines for English-Dutch Translation Quality Assessment, version 1.0*. LT3 Technical Report - LT3 13.02. available from lt3.hogent.be/en/publications/annotation-guidelines-for-english-dutch-translation-quality-/

EN 15038. 2006. *Translation services - Service requirements*

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

Localization Industry Standards Association. LISA QA Model 3.1. available from www.lisa.org/LISA-QA-Model-3-1.124.0.html

Macken, Lieve, Orphée De Clercq, and Hans Paulussen. 2011. *Dutch Parallel Corpus: a Balanced Copyright-Cleared Parallel Corpus*, Meta, 56(2): 374-390. Les Presses de l'Université de Montréal.

O'Brien, Sharon. 2012. *Towards a Dynamic Quality Evaluation Model for Translation*, The Journal of Specialised Translation(17):55-77.

Plitt, Mirko and François Masselot. 2010. *Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context.*, The Prague Bulletin of Mathematical Linguistics, 93:7-16.

Rinsche, Adriane and Nadia Portera-Zanotti. 2009. *The size of the language industry in the EU.* Retrieved from http://ec.europa.eu/dgs/translation/publications/studies/index_en.htm

SAE J2540. December 2001. *Quality Metric for Language Translation.* www.apex-translations.com/documents/sae_j2450.pdf

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. *EACL 2012, Proceedings of the Demonstrations Session at the 13th European Chapter of the Association for Computational Linguistics.*, Avignon, France.

Tatsumi, Midori. 2010. *Post-Editing Machine Translated Text in a Commercial Setting: Observation and Statistical Analysis.* Dublin: Dublin City University.

Toury, Gideon. 1995. *The Nature and Role of Norms in Translation*, Descriptive Translation Studies and Beyond:53-69. Amsterdam-Philadelphia: John Benjamins.

Van Slype, Georges. 1979. *Critical Methods for Evaluating the Quality of Machine Translation. Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management*, Report BR-19142. Bureau Marcel van Dijk.

White, John. 1995. Approaches to Black-box Machine Translation Evaluation. *Proceedings of the MT Summit 1995*, Luxembourg.

Williams, Malcolm. 2009. *Translation Quality Assessment*, Mutatis Mutandis 2(1):3-23.