Two Sides of the Same Coin: Assessing Translation Quality in Two Steps through Adequacy and Acceptability Error Analysis

Joke Daems, Lieve Macken, Sonia Vandepitte
Department of Translation, Interpreting and Communication, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
E-mail: joke.daems@ugent.be, lieve.macken@ugent.be, sonia.vandepitte@ugent.be

A translator has to find the balance between adhering to the norms of the source text (adequacy) and respecting the norms of the target text (acceptability) (Toury, 1995). The quality of a translation can then be judged on its (non-)adherence to these norms. This is a common quality judgment for machine translation, where evaluators give translated segments an adequacy and acceptability (sometimes 'fluency') score on a scale from one to five (White, 1995).

When looking at translation quality assessment through error analysis, however, the dichotomy between acceptability and adequacy is not always as distinct. Existing metrics do provide error categories relating to both types of issues. For example, QTLaunchPad's MQM has a category for fluency and one for accuracy; TAUS suggests a category for accuracy, one for terminology, and two categories that could relate to acceptability (language and style); FEMTI proposes suitability, accuracy and wellformedness; and MeLLANGE offers categories for language and content transfer. Yet these categories are all part of one and the same evaluation step: evaluators have to identify issues and assign the correct category to these issues. Research has shown that deciding whether an error belongs to adequacy or acceptability is one of the most difficult aspects of error analysis for human annotators, together with having to assign an error weight to each error instance (Stymne & Ahrenberg, 2012).

We therefore propose facilitating the error annotation task by introducing an annotation process which consists of two separate steps that are similar to the ones required in the European Standard for translation companies EN 15038: an error analysis for errors relating to acceptability (where the target text as a whole is taken into account, as well as the target text in context), and one for errors relating to adequacy (where source segments are compared to target segments). We present a fine-grained error taxonomy suitable for a diagnostic and comparative analysis of machine translated-texts, post-edited texts and human translations. Categories missing in existing metrics have been added, such as lexical issues, coherence issues, and text type-specific issues. Annotator subjectivity is reduced by assigning error weights to each error category beforehand, which can be tailored to suit different evaluation goals, and by introducing a consolidation step, where annotators discuss each other's annotations.

The approach has been tested during two pilot studies with student translators who both post-edited and translated different texts. Inter-annotator agreement shows that the proposed categorization is clear and that it is necessary to include a consolidation phase. Annotations after consolidation were used to analyze the most common errors for each method of translation and to provide an average error score per word for each text. In a next phase, the annotations were manually grouped into source text-related error sets: a source text passage and the translations for that passage that contain errors. Error sets allow for a diagnostic evaluation: which source text-segments that were problematic for machine translation are still problematic after post-editing and how? How many and which post-editing errors originate from the machine translation output?

Though the approach in its current form requires much time and human effort (the annotation process in itself costs around 45 minutes for 150 words of a new MT text, with acceptability annotations requiring the most time: 30 minutes), it does provide rich data needed to improve translation quality. Familiarity with a text can seriously decrease annotation time, and the time for HT or PE if also lower than for MT. We are currently optimizing the annotation process to increase the speed and reduce manual effort, and we believe that the processing of the annotations and the creation of the error sets can, at least in part, be automated.