

Assessing Inter-Annotator Agreement for Translation Error Annotation

Arle Lommel, Maja Popović, Aljoscha Burchardt,

DFKI

Alt-Moabit 91c, 10559 Berlin, Germany

E-mail: arle.lommel@dfki.de, maja.popovic@dfki.de, aljoscha.burchardt@dfki.de

Abstract

One of the key requirements for demonstrating the validity and reliability of an assessment method is that annotators be able to apply it consistently. Automatic measures such as BLEU traditionally used to assess the quality of machine translation gain reliability by using human-generated reference translations under the assumption that mechanical similarity to references is a valid measure of translation quality. Our experience with using detailed, in-line human-generated quality annotations as part of the QTLaunchPad project, however, shows that inter-annotator agreement (IAA) is relatively low, in part because humans differ in their understanding of quality problems, their causes, and the ways to fix them. This paper explores some of the facts that contribute to low IAA and suggests that these problems, rather than being a product of the specific annotation task, are likely to be endemic (although covert) in quality evaluation for both machine and human translation. Thus disagreement between annotators can help provide insight into how quality is understood.

Our examination found a number of factors that impact human identification and classification of errors. Particularly salient among these issues were: (1) disagreement as to the precise spans that contain an error; (2) errors whose categorization is unclear or ambiguous (i.e., ones where more than one issue type may apply), including those that can be described at different levels in the taxonomy of error classes used; (3) differences of opinion about whether something is or is not an error or how severe it is. These problems have helped us gain insight into how humans approach the error annotation process and have now resulted in changes to the instructions for annotators and the inclusion of improved decision-making tools with those instructions. Despite these improvements, however, we anticipate that issues resulting in disagreement between annotators will remain and are inherent in the quality assessment task.

Keywords: translation, quality, inter-annotator agreement

1. Introduction

The development and improvement of Machine Translation (MT) systems today makes heavy use of human knowledge and judgments about translation quality. Human insight is typically provided in one of four ways:

1. human-generated reference translations
2. rating of MT output based on perceived quality
3. post-edits of MT output (implicit error markup)
4. explicit error markup of MT output.

However, it is well known that human judgments of translation show a high degree of variance: in WMT testing, the inter-annotator agreement (IAA), i.e., agreement between two or more annotators, in a rating task did not exceed 0.40 (α , described in section 3) and *intra*-annotator agreement (i.e., the agreement of raters *with themselves* when faced with the same assessment task multiple times) did not exceed 0.65 (Bojar et al., 2013:6–8). By contrast, for most IAA tasks, agreement of at least 0.85 is required for a measure to be considered reliable.

It must be put forth as a fundamental assumption that there is no single, objectively “correct” translation for a given text, but rather a range of possible translations that range from perfectly acceptable to totally unacceptable. Moreover, Translation quality is always relative to given specifications or the given job. Factors like resource availability, production environment, target audience, etc. can determine whether a certain translation is considered correct or not. For example, in an on-demand instant MT system, quality may be determined by

whether or not the text enables the reader to accomplish a task. In such cases texts may show low levels of Accuracy and Fluency and yet still be considered to meet quality expectations.¹ Although we will not discuss this issue in depth in this paper, it should be kept in mind.

The realization that there is a spectrum of acceptable translations rather than a single optimal output and that raters will often disagree in their opinions are reasons why automatic measures of MT quality like BLEU have been designed to be able compare MT output with multiple human translation references from the very beginning (Papimemi et al. 2002).

Considering the four types of human insight listed at the start of this paper, the question of inter-annotator agreement boils down, in part to questions such as: How similar are two or more human reference translations? How similar are ratings? How similar are post-edits? How similar are explicit error markups? In all of these cases, any subsequent experiments using performance measures like BLEU or METEOR or analysis tools like Hjerson (Popović 2011) rely on the assumption that the human input provides a reliable basis.

To the best of our knowledge, the question of how many reference translations, ratings, or post-edits are needed per sentence to substantiate reliable and replicable

¹ As a result of this realization, there has been a recent shift towards the use of explicit specifications that guide translation, assessment, and postediting (Melby, Fields, & Housley, 2014).

quality judgments about MT performance has not yet found a widely accepted answer. In this paper, we will report first steps in evaluating inter-annotator agreement for the case of explicit error markup.

As MT errors can overlap or interact in many ways, we will focus on machine translations that show only few errors to minimize the problem of overlapping errors.

One reason for human disagreement in the case of analysis based on post-edits or manual error annotation is the simple fact that errors can often be analyzed (or explained) in multiple ways. For example, a seemingly missing plural *-s* in an English noun phrase might constitute an *agreement* error (*Fluency*) or indicate a *mistranslation* of a noun, which was meant to be singular (*Accuracy*). When translating from Chinese, for example, such factors may lead to different opinions of human translators since Chinese does not mark number; such confusion is likely inherent in the task since there are multiple valid ways to understand an error.

The remainder of this paper will focus on some of the issues that complicate the determination of IAA with examples from a human annotation campaign undertaken by the QTLaunchPad project.

2. Experimental setup

In the annotations described in this paper multiple professional translators from commercial language service providers (LSPs) were asked to evaluate a set of 150 sentences in one of four language pairs (EN>ES, ES>EN, EN>DE, and DE>EN) using the open-source translate5 (<http://www.translate5.net>) tool.

The sentences were selected from the WMT 2012 shared task data produced by state-of-the-art MT systems. The sentences were selected so that only those with a “native” source were used (i.e., only those sentences where the source segment had been written in the source language rather than translated from another language).² To select the sentences for annotation, human evaluators reviewed the MT output for the 500 translations of each of the systems—SMT, RbMT, and (for English source only) hybrid—plus the 500 reference human translations. These reviewers ranked each translated segment according to the following scale:

- Rank 1: Perfect output (no edits needed)
- Rank 2: “Near misses” (1–3 edits needed to be acceptable)
- Rank 3: “Bad” (>3 edits needed)

From the Rank 2 sentences, we pseudo-randomly selected a corpus of 150 sentences, to create the “calibration set.” The calibration set consisted of the following breakdown of segments by production type:

- EN>ES and EN>DE: 40 segments each SMT, RbMT, and hybrid, plus 40 human translations.
- ES>EN and DE>EN: 60 segments each SMT and RbMT, plus 40 human translations.

These corpora were uploaded into the translate5 system and the annotators were all provided with a set of written guidelines³ and invited to attend or view a recording of a webinar⁴ introducing them to the tool and task.

The segments were annotated by three (DE>EN), four (EN>ES, ES>EN), or five (EN>DE) annotators. Annotators were encouraged to interact with our team and to ask questions. The annotators used translate5 to associate issues with specific spans in target sentences. The list of issues used was the following:

- Accuracy
 - Terminology
 - Mistranslation
 - Omission
 - Addition
 - Untranslated
- Fluency
 - Register/Style
 - Spelling
 - Capitalization
 - Typography
 - Punctuation
 - Grammar
 - Morphology (word form)
 - Part of speech
 - Agreement
 - Word order
 - Function words
 - Tense/mood/aspect
 - Unintelligible

The definitions for each of these issues are provided in the downloadable guidelines previously mentioned. Annotators were instructed to select “minimal” spans (i.e., the shortest span that contains the issue) and to add comments to explain their choices, where relevant.

Annotators found the numbers of issues given in Table 1.

| | ES>EN | EN>ES | DE>EN | EN>DE | All |
|----------|-------|-------|-------|-------|------|
| Annot. 1 | 157 | 387 | 219 | 216 | — |
| Annot. 2 | 229 | 281 | 266 | 278 | — |
| Annot. 3 | 98 | 289 | 327 | 277 | — |
| Annot. 4 | 255 | 235 | — | 315 | — |
| Annot. 5 | — | — | — | 278 | — |
| TOTAL | 739 | 1192 | 812 | 1364 | 4107 |
| AVG | 185 | 298 | 271 | 273 | 257 |
| AVG/Seg | 1.23 | 1.99 | 1.80 | 1.82 | 1.71 |

Table 1: Number of issues found in corpus per annotator and language pair.

² WMT data includes both sentences written in the source language and those translated into the source language from another language.

³ <http://www.qt21.eu/downloads/shared-task-webinar.mov>

⁴ <http://www.qt21.eu/downloads/webinar-on-shared-task.pdf>

The distribution of identified issues in this corpus is described in Burchardt et al. (2014) and is not covered here, as the analysis of specific issue types and their distribution is beyond the scope of this paper.

3. Assessing Inter-Annotator Agreement

As part of the evaluation of the results of the annotation task, we wished to determine inter-annotator agreement (IAA), sometimes known as inter-rater reliability. Demonstrating a high degree of IAA is a necessary step to showing that an assessment metric is reliable. In addition, demonstrating reliability helps, but is not sufficient, to demonstrate that a metric is fair.

There are a number of different approaches to demonstrating IAA. One approach is to look at absolute agreement between raters. This approach typically overstates agreement, however, because it does not take into account the probability of agreement by chance. For example, if items are assessed on a 1 to 5 scale with an equal distribution between each of the points on the scale, an assessment that *randomly* assigns scores to each item would achieve an absolute agreement approaching 0.2 (i.e., 20% of numbers would agree) as the sample size approaches infinity. As a result, for many tasks a different measure, Cohen's kappa (κ)⁵ is preferable because it attempts to take the probability of random agreement into account, although the assumption that annotators will make random choices in the absence of a clear option is debatable, a point we will return to, so κ scores may *understate* agreement (Uebersax, 1987). Nevertheless, in order to provide comparison with assessments of IAA given in WMT results, this study uses κ scores.

To calculate scores, we examined the positional tagging for issues in pairwise comparisons between each annotator, averaging the results within each language pair. Figure 1 shows an example in which one annotator tagged two issues and the second tagged one. In the last row the lighter cells show the area of disagreement.

Source: While she's a U.S. citizen, she sees herself as a part of two countries.

| | | | | | | | | | | | | | | |
|-----------------------|----------------|---------|-----|-----|------------|-----|-------|------|--------|-----|------|-----|--------|---------|
| | | Während | sie | ein | US-Bürger, | sie | sieht | sich | selbst | als | Teil | der | beiden | Länder. |
| A1 | Mistranslation | | | | | | | | | | | | | |
| | Word Order | | | | | | | | | | | | | |
| A2 | Mistranslation | | | | | | | | | | | | | |
| | Word order | | | | | | | | | | | | | |
| A1 / A2 = .85 (11/13) | | | | | | | | | | | | | | |

Figure 1. IAA for an English>German translation (absolute agreement average = .85, Kappa IAA = .72)

⁵ Kappa is calculated as follows:

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)}$$

where P(a) is probability of actual agreement, i.e., $\sum_k p(a1 = k, a2 = k)$ and P(e) is probability of agreement by chance, i.e., $\sum_k p(a1 = k) * p(a2 = k)$, where k denotes class (in this case the error tag) and a1, a2 refer to the two annotators.

Because κ is appropriate only for pair-wise comparisons, we evaluated the similarity between each pair of annotators separately and took the average score, as shown in Figure 2. In this example three different IAA figures are assessed, one for each of the three possible pair-wise comparisons. In this example, Rater 1 and Rater 3 are quite similar with $\kappa = 0.89$ while Rater 2 differs from both of them with $\kappa = 0.57$ with Rater 1 and $\kappa = 0.53$ with Rater 3. Although both Rater 2 and Rater 3 identified the same types of errors (and were alike in not identifying the Agreement error identified by Rater 1), they disagreed on the precise spans for those errors, leading to lower κ scores.

Source: A first year PPE student, who, ironically, had been to Eton, said: "You're the daughter of a fascist pig."

| | | | | | | | | | | | | | | | | | | | | | |
|-----------------------|----------------|----|--------|-----|------------|----|------|------|---------------|-------|------|---|-------|-------|-----|------|----|----|-------|------------|--|
| | | Un | primer | año | estudiante | de | PPE, | que, | irónicamente, | había | sido | a | Eton, | dijo: | "Es | hija | de | un | cerdo | fascista." | |
| A1 | Word order | | | | | | | | | | | | | | | | | | | | |
| | Mistranslation | | | | | | | | | | | | | | | | | | | | |
| | Agreement | | | | | | | | | | | | | | | | | | | | |
| A2 | Word order | | | | | | | | | | | | | | | | | | | | |
| | Mistranslation | | | | | | | | | | | | | | | | | | | | |
| | Agreement | | | | | | | | | | | | | | | | | | | | |
| A3 | Word order | | | | | | | | | | | | | | | | | | | | |
| | Mistranslation | | | | | | | | | | | | | | | | | | | | |
| | Agreement | | | | | | | | | | | | | | | | | | | | |
| A1 / A2 = .79 (15/19) | | | | | | | | | | | | | | | | | | | | | |
| A1 / A3 = .95 (18/19) | | | | | | | | | | | | | | | | | | | | | |
| A2 / A3 = .84 (16/19) | | | | | | | | | | | | | | | | | | | | | |

Figure 2. IAA for an English>Spanish translation (absolute agreement average = .86, Kappa IAA = .66)

Note that if a simpler segment-level measure that counts only whether the same issue classes were identified for each segment were used instead, the results would be rather different. In that case the example in Figure 1 would yield an agreement figure of 0.5 (there would be a total of two issues for the segment and the annotators would agree on one). For the example in Figure 2, by contrast, Rater 1 would show the same agreement with Raters 2 and 3 (.67) while Rater 2 and Rater 3 would show perfect agreement (1.0) since they identified the same issues, even though they disagreed on the scope.

Using kappa allowed us to calculate κ scores for the test data sets, as shown in Table 2. (The EN>DE pair was annotated by five reviewers, but one was received after this analysis was completed.) The results of this analysis lie between 0.18 and 0.36 and are considered to be "fair" (see Bojar et al., 2013:6–8, for discussion of κ levels). The overall average is 0.30.

| | ES>EN | EN>ES | DE>EN | EN>DE |
|---------|-------|-------|-------|-------|
| a1-a2 | 0.30 | 0.35 | 0.23 | 0.36 |
| a1-a3 | 0.18 | 0.36 | 0.36 | 0.28 |
| a2-a3 | 0.19 | 0.28 | 0.29 | 0.33 |
| a1-a4 | 0.25 | 0.33 | | 0.30 |
| a2-a4 | 0.26 | 0.36 | | 0.34 |
| a3-a4 | 0.34 | 0.35 | | 0.30 |
| Average | 0.25 | 0.34 | 0.29 | 0.32 |

Table 2. Kappa coefficients measuring inter-annotator agreement for MQM error annotation

By comparison, the WMT organizers evaluated κ scores for their rating tasks in which raters were asked to assign quality rates from 1–4 (2011/2012) or 1–5 (2013). The κ scores for this task are presented in Table 3.

| | ES>EN | EN>ES | DE>EN | EN>DE |
|----------|-------|-------|-------|-------|
| WMT 2011 | 0.38 | 0.37 | 0.32 | 0.49 |
| WMT 2012 | 0.36 | 0.25 | 0.38 | 0.30 |
| WMT 2013 | 0.46 | 0.33 | 0.44 | 0.42 |
| Average | 0.40 | 0.32 | 0.38 | 0.40 |

Table 3. κ scores for WMT ranking tasks.

As can be seen, the IAA scores for the human annotation task are lower than those for the rating task, with the highest scores in the annotation task roughly on par with the lowest scores in the rating task. While such a result might seem discouraging, we believe there are a number of reasons for this result and that our seemingly low results may reveal problems hidden in many translation quality assessment tasks/methods. The remainder of this paper will address some of these results.

4. Scope of span-level annotation

One fundamental issue that the QTLaunchPad annotation encountered was disagreement about the precise scope of errors. In the example shown in Figure 2, for instance, Annotator 1 marked the following issue spans:

Un **primer año estudiante** de PPE, que, irónicamente, había **sido** a Eton, dijo: “**Es** hija de un cerdo fascista”.

while Annotator 2 marked the following spans:

Un **primer año estudiante de PPE**, que, irónicamente, **había sido** a Eton, dijo: “Es hija de un cerdo fascista”.

Here they fundamentally agree on two issues (a *Word order* and a *Mistranslation*) and disagree on the third (an *Agreement*). However, for the two issues they agree on, they disagree on the span that they cover. Annotators were asked to mark *minimal* spans, i.e., spans that covered *only* the issue in question, but they frequently disagreed as to what the scope of these issues was.

In the case of *primer año estudiante* vs. *primer año estudiante de PPE*, two word orders are equally acceptable: *estudiante de primer año de PPE* and *estudiante de PPE de primer año*. Thus it seems that the reviewers agreed that the phrase (*Un*) *primer año estudiante de PPE* was problematic, but disagreed as to the solution and whether *de PPE* needed to be moved or not.

In the case of *sido* vs. *había sido*, the correct rendering would be *había ido* (‘had gone’) instead of *había sido* ‘had been’. Annotator 1 thus correctly annotated the minimal span, while Annotator 2 annotated a longer span. However, it may be that the two reviewers perceived the issue differently and that the cognitively relevant span for Annotator 1 was the word *sido* while for Annotator 2 it was the entire verbal unit, *había sido*.

In these two cases we see that reviewers can agree on the nature (and categorization) of issues and yet still disagree on their precise span-level location. In some instances this disagreement may reflect differing ideas about optimal solutions, as in the case of whether to include *de PPE* in the *Word order* error. In others the problem may have more to do with perceptual units in the text.

In such cases we are uncertain how best to assess IAA. Using the model presented in the previous section these are marked as agreement for some words and disagreement for others. The net effect is that, at the sentence level, they have partial agreement.

| | ES>EN | EN>ES | DE>EN | EN>DE |
|-------------------|-------|-------|-------|-------|
| Accuracy | 0.0% | 0.1% | 0.0% | 0.4% |
| Addition | 0.5% | 1.3% | 0.4% | 2.2% |
| Agreement | 0.4% | 2.8% | 0.3% | 1.4% |
| Capitalization | 0.0% | 0.6% | 0.3% | 0.3% |
| Fluency | 0.0% | 0.0% | 0.0% | 0.0% |
| Function words | 9.2% | 10.1% | 4.1% | 1.9% |
| Grammar | 3.0% | 0.3% | 0.1% | 9.5% |
| Mistranslation | 6.4% | 6.9% | 4.4% | 8.0% |
| Morphology | 0.0% | 0.1% | 1.0% | 0.1% |
| POS | 1.1% | 0.5% | 1.2% | 0.0% |
| Punctuation | 2.0% | 0.7% | 1.2% | 1.5% |
| Spelling | 0.4% | 0.6% | 0.1% | 0.2% |
| Style/Register | 7.1% | 7.4% | 3.8% | 6.3% |
| Tense/Aspect/Mood | 1.6% | 4.4% | 0.5% | 2.3% |
| Terminology | 6.3% | 14.2% | 8.9% | 2.8% |
| Typography | 0.0% | 0.4% | 0.0% | 0.0% |
| Unintelligible | 0.1% | 0.0% | 1.7% | 1.2% |
| Untranslated | 0.3% | 0.0% | 0.3% | 0.5% |
| Word order | 8.0% | 10.1% | 24.2% | 6.1% |

Table 4. Percentage of instances for each issue class in which annotators disagreed on precise spans.

Quantitatively, the impact of different assessments of spans can be seen in Table 4, which shows, based on a pairwise comparison of annotators, the percentage of cases in which annotators differ in their assessment of the location (but not the nature) of spans. Note that this analysis does not distinguish between cases where spans are actually related and where they are independent instances of the same category (e.g., two annotators annotate totally different *Mistranslations* in a segment), so it may overstate the numbers slightly.

In this case it can be seen that *Word order* has the highest overall rate of instances in which annotators disagreed on the precise location of spans. From other analysis done in the QTLaunchPad project we know that word order is particularly problematic for German>English translations, and here we see a high confusion rate for this issue type. It is not surprising that *Word order* ranks so highly in terms of confusion because often there are different ways to interpret ordering errors. So even though annotators largely agree on the existence of the problem, they often disagree on the location.

5. Unclear error categorization

In the example discussed in the last section, one item was tagged by Annotator 1 and missed by other reviewers. Annotator 1 tagged it as *Agreement*, but a close examination of the issue leaves it unclear why *Agreement* was chosen. The use of *Es* is clearly a mistake since it cannot generally mean “You’re”. After consulting with a Spanish native speaker, it appears that the error should definitely have been tagged (so two of three reviewers missed the problem) but that there are multiple possible categorizations depending on how *You’re* should be rendered in Spanish. Possible options include the following:

- **Mistranslation.** *Es* ‘is’ clearly not the intended meaning. *Es* can thus be treated as a mistranslation for *Tu eres* or *Eres* ‘You (informal) are’.
- **Omission.** If a formal register is intended (an unlikely choice for a human translator in this case, but possible since MT systems might be optimized to use the formal), then *Usted es* would be the appropriate text, and there would be an *Omission* of *Usted*.
- **Agreement.** Since Spanish can, in most circumstances, drop subject pronouns (although, generally, *Usted* should not be dropped), *Es* could exhibit an *Agreement* problem with the implicit subject *tu*.

Since the exact nature of the problem is not clear from the text (source or target), the rules used in the annotation task would specify that the first possible one in the list of issues be taken, unless a more specific type also applies. In this case, then, *Mistranslation* would be the appropriate issue type. However, if the annotator did not perceive the phrase as having the wrong meaning, but rather an awkward phrasing, then the annotator might never arrive at this option.

The problem of differing assessments of the nature of problems is pervasive in our corpus, as shown in Table 5, which provides the percentage of times in which one annotator marked a sentence as having a specific issue and another annotator did not mark that same issue type as occurring within the sentence.

The figures in Table 5 were derived in pair-wise comparisons between annotators. For each case if one annotator noted a specific class of issue, regardless of location within the segment, and another annotator also annotated the same issue class as occurring in the segment, the annotators were deemed to be in agreement. If one annotator noted an issue class and another did not then they were deemed to be in disagreement. This provides a rough measure for the frequency with which the issues might be annotated in different ways. Examining the totals reveals the following notable points:

- *Mistranslation* and *Terminology* show high levels of confusion. (Burchardt et al. 2014 discusses the confusion between these two and the correlation with the length of the

| | ES>EN | EN>ES | DE>EN | EN>DE |
|-------------------|-------|-------|-------|-------|
| Accuracy | 0.2% | 0.2% | 0% | 1.0% |
| Addition | 2.1% | 4.8% | 4.0% | 3.5% |
| Agreement | 6.2% | 7.3% | 3.6% | 4.7% |
| Capitalization | 0.3% | 2.9% | 1.1% | 1.2% |
| Fluency | 0% | 3.0% | 0% | 0.2% |
| Function words | 30.4% | 21.9% | 18.9% | 7.6% |
| Grammar | 6.3% | 1.0% | 0.7% | 16.8% |
| Mistranslation | 23.6% | 22.8% | 27.1% | 24.1% |
| Morphology | 0.2% | 0.3% | 3.8% | 5.4% |
| Omission | 5.3% | 6.6% | 7.6% | 5.2% |
| POS | 2.9% | 2.2% | 2.4% | 1.0% |
| Punctuation | 4.0% | 4.5% | 9.1% | 9.3% |
| Spelling | 0.8% | 2.2% | 1.1% | 0.9% |
| Style/Register | 16.3% | 9.1% | 3.3% | 11.0% |
| Tense/Aspect/Mood | 3.9% | 11.3% | 3.1% | 7.1% |
| Terminology | 12.9% | 24.5% | 19.1% | 13.1% |
| Typography | 0.2% | 0.8% | 0.2% | 0.0% |
| Unintelligible | 0.2% | 0.0% | 1.3% | 1.2% |
| Untranslated | 0.9% | 0.9% | 0.9% | 0.5% |
| Word order | 7.2% | 5.9% | 8.9% | 4.4% |
| No error | 15.4% | 10.4% | 7.6% | 8.7% |

Table 5. Percentage of instances at the sentence level in which one annotator noted an issue and another annotator did not, by language pair.

problematic span, with *Mistranslation* being used for longer spans in general while *Terminology* is used primarily for single-word spans.)

- The *Function word* category also shows very high confusion, with very different profiles between the language pairs. Overall, this category was one of the most frequently occurring and problematic in the entire corpus.
- *Word order* shows high levels of agreement between annotators, although the span-level agreement is significantly lower.
- There is a relatively high percentage of sentences where some annotators say that there are no errors and other annotators say there are some errors.

5.1. Confusion within the hierarchy

It is important to note that the MQM issues exist in a hierarchy and the annotators were instructed that if no issue applied precisely at one level in the hierarchy, they should select the next highest level. As a result, annotators may be confused about which class applies to a specific error and find the issue types confusing. When we ran the annotation campaign a number of annotators came back to us with cases where they were unsure as to which level was appropriate for a given issue.

For example, if the annotator encountered a grammatical error but none of the children of *Grammar* applied (e.g., a sentence has a phrase like “he slept the baby” in which an intransitive *very* is used transitively, but there is no precise category for this error, which is known as a *Valency* error), then the parent (*Grammar* should be used). As a result, many issues could be annotated at

multiple levels in the hierarchy, especially if the precise nature of the error is not entirely clear, as with the example given above, where it *could* be *Agreement*, but is not clearly so. In such cases some annotators may pick a specific category, especially if they feel comfortable with the category, while others may take the more general category in order to be safe in a situation where they are not certain.

5.2. Lack of clear decision tools

One of the problems annotators faced was the mismatch in knowledge between the team that created the MQM metric and themselves. Many of the training materials we created assumed a certain degree of background knowledge in linguistics that it turned out we could not assume. A simple list of issue types and definitions along with some general guidelines were insufficient to guide annotators when faced with unfamiliar issues which they intuitively know how to fix but which they are not use to classifying in an analytic scheme.

As a result, we discovered that annotators need better decision-making tools to guide them in selecting issues, especially when they are easily confusable, as is the case with issues in the hierarchy, or when there are multiple, equally plausible explanations for an error. By formalizing and proceduralizing the decision-making process, confusion could be reduced.

6. Annotators' personal opinions

Finally, we cannot discount the possibility that different translators may simply disagree as to whether something constitutes an error or not, based on dialect, idelect, education, or even personal opinion. Such cases, where one speaker of a language sees a sentence as acceptable and another does not, have long been the bane of linguistics professors who want to have a clear-cut case for putting a star (*) on unacceptable sentences. In addition, although we provided the annotators with detailed guidelines for the issue types, they may have disagreed as to whether something was serious enough to annotate. Thus the individual annotators' opinions are likely to have a substantial impact on overall IAA, albeit one hard to quantify without an extensive qualitative consultation with annotators in a lab setting.

7. Lessons learned and conclusions

Analytic measures like MQM offer the potential to gain insights into the causes of translation problems and how to resolve them. Although IAA in our first studies reported here is lower than ideal, we believe that our findings point out a covert problem in most annotation and quality evaluation tasks. As we discovered, the human annotators' meta-understanding of language is quite variable, even when working with professional translators. Even with an analytic framework and guidelines there is significant, and perhaps unavoidable, disagreement between annotators. To a large extent this disagreement reflects the variability of human language.

Evaluation methods that rely on reference translations such as METEOR or BLEU must assume that the range of available translations provides a "good enough" approximation of the range of language variation that can

be expected in translations. This assumption may be valid for limited cases in which the training data used for MT is substantially similar to the reference translations, but in cases with heterogeneous training data and references, it is entirely possible that reference-based methods may penalize acceptable translations because they differ from references and reward less optimal answers because they are mechanically similar to references.

In order to improve future MQM assessment and improve IAA rates, we have revised the issue hierarchy to reduce certain distinctions (e.g., between *Typography* and *Punctuation*) that offered little discriminatory power, while we have added more details to other categories such as splitting the *Function words* category to allow more detailed analysis of specific problems. We have also created a formal decision tree and improved guidelines⁶ to assist with future annotation work and to help annotators distinguish between problematic categories.

While improved IAA is an important goal where possible, the exact nature of disagreement where clarification of issue types and procedures does not result in agreement can also provide insight into how humans conceive of translation quality. If one of the goals of MT research is to deliver translation closer to human quality, a better understanding of the variables that impact quality judgments is vital, as is understanding the extent of variation that comprises "acceptable" translation. This study and the issues it raises will help provide better understanding of these factors. We intend to continue this analysis using our improved issue hierarchy and decision tools in an annotation campaign planned for March 2014.

8. Acknowledgements

This study was conducted as part of the European Commission-funded project QTLAUNCHPAD (Seventh Framework Programme (FP7), grant number: 296347). The authors thank Ergun Biciçi (Dublin City University) for assistance with calculating IAA scores. Thanks also to Nieves Sande (DFKI) for her kind assistance with Spanish-language issues.

9. References

- Bojar, O.; Buck, C.; Callison-Burch, C.; Federmann, C.; Haddow, B.; Koehn, P.; Monz, C.; Post, M.; Soricut, R. and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation, pp 1–44. (<http://www.statmt.org/wmt13/pdf/WMT01.pdf>.)
- Burchardt, A.; Gaspari, F.; Lommel, A.; Popović, M. and Toral, A. (2014). Barriers for High-Quality Machine Translation (QTLAUNCHPAD Deliverable 1.3.1). http://www.qt21.eu/launchpad/system/files/deliverable_s/QTLP-Deliverable-1_3_1.pdf.
- Melby, A.; Fields, P.J. and Housley, J. (forthcoming). Assessment of Post-Editing via Structured Translation Specifications To appear in M. Carl, S. O'Brien, M.

⁶ <http://www.qt21.eu/downloads/annotatorsGuidelinesNew.pdf>

- Simard, L. Specia, & L. Winther-Balling (Eds.), *Post-editing of Machine Translation: Processes and Applications*. Cambridge: Cambridge Scholars Publishing.
- Papineni, K.; Roukos, S.; Ward, T. and Zhu, W.J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, pp. 311–18.
- Popović, M. (2011). Hjerion: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *Prague Bulletin of Mathematical Linguistics* 96, pp. 59–68.
- Uebersax, J.S. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin* 101, 140–46.