Natural Language Processing (2025), **31**, pp. 1017-1041
doi:10.1017/nlp.2024.5

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Calibration and context in human evaluation of machine translation

Rebecca Knowles and Chi-kiu Lo

National Research Council of Canada, Ottawa, ON, Canada
Corresponding author: Rebecca Knowles; Email: Rebecca.Knowles@nrc-cnrc.gc.ca

(Received 12 April 2023; revised 27 October 2023; accepted 1 December 2023; first published online 3 June 2024)

Special Issue on '**The Role of Context in Neural Machine Translation Systems and its Evaluation**', guest-edited by Rebecca Knowles and Sheila Castilho

Abstract

Human evaluation of machine translation is considered the "gold standard" for evaluation, but it remains a challenging task for which to define best practices. Recent work has focused on incorporating intersentential context into human evaluation, to better distinguish between high-performing machine translation systems and human translations. In this work, we examine several ways that such context influences evaluation and evaluation protocols. We take a close look at annotator variation through the lens of calibration sets and focus on the implications for context-aware evaluation protocols. We then demonstrate one way in which degraded target-side intersentential context can influence annotator scores of individual sentences, a finding that supports the context-aware approach to evaluation and which also has implications for best practices in evaluation protocols.

Keywords: machine translation; evaluation

1. Introduction

Human evaluation of machine translation is considered the "gold standard" (as compared to automatic evaluation), but ensuring that this gold standard is of high quality is a difficult task. A recent challenge in human annotation of machine translation has been determining how best to incorporate context. In this work, we take a close look at several aspects of context-informed human annotation of machine translation quality. We use data from recent WMT (Conference on Machine Translation) human evaluations as the basis for our analyses. Throughout this paper, the main type of context examined consists of the preceding sentences from the same document, either on the source side or on the target side (machine translated output), a type of intersentential context.[a]

There are numerous sources of potential human error or variation in annotation. Across annotators, there can be differences due to annotator strictness or leniency. Differences in annotation behavior could be related to underspecification of the task description: for example, if there isn't

[a]Using the five types of context described in Melby and Foster (2010), this might be best described as a subset of the co-text (in our setting, the source language text preceding the sentence being evaluated) and either bi-text or the co-text of the target translation (for the target segments of the same document that were observed and scored prior to scoring the segment in question).

enough context to determine whether a translation is correct or merely has the potential to be correct (such as when the target language requires types of information like number, gender, tense, etc. that are not provided by the source), should annotators score this as a perfect translation or not? There could also be sources of variation across annotators that are not errors, for example, what is an ideal translation in one dialect may be erroneous in another. A single annotator's scores may also not always be perfectly self-consistent. In annotation protocols that utilize sliders to assign scores to translations, there could be physical explanations for variation: if the slider is very sensitive, a slight mouse movement could result in relatively large score differences. There could also be sources of genuine error: the annotator could be in a hurry, they could misread something, they might be focused on a different aspect of quality each time they annotate the segment, or they could fail to realize that there is not enough context to definitively give a specific quality score. Annotator error and variation may also change over time, as the annotator becomes more familiar with the task or more fatigued.

The challenge of reliable annotation is often described as a problem of dealing with human annotator inconsistency. However, close analyses of annotation protocols have repeatedly brought to light sources of error that come from the protocols themselves, rather than the human annotators (or, in some cases, the interplay between the two). Using calibration data from WMT 2022 (where multiple annotators annotated the same set of data, assigning scores between 0 and 100 to each segment), we are able to consider how different annotators use the scale of possible scores. We show that small features of the interface can influence annotator behavior and also raise questions about whether additional instruction for annotators could produce different results. We also use this data to reevaluate questions of annotator consistency and correlation, as well as to consider strengths and weaknesses of approaches to the standardization of annotator scores.

The introduction of context into evaluation has broad consequences, from interface design to annotation protocols to standardization of scores, and it is important to revisit assumptions and approaches in light of these changes, in order to build annotation frameworks that are as reliable as possible. A consequence of our areas of focus and the use of existing WMT data is that we are able to draw some conclusions about ways of improving this particular approach to human annotation, but to draw broad and reliable conclusions about what types of annotation data to collect or what styles of annotation produce the most precise annotations would require additional experimentation and data collection.

To further understand the influence of intersentential context on evaluation, we are able to make use of quality control items from WMT shared tasks to examine what happens when annotators score identical sentences with the source context held constant while the target intersentential context is modified, specifically the case in which the quality of some target sentences is degraded. This enables us to shed additional light on how segment-level scores change based on target intersentential context, as well as to identify best practices in data collection and quality control. In contrast to prior work that has typically focused on varying the number of preceding sentences shown, this work focuses on variations in the quality of those translations, compared to one another in settings where the number of preceding sentences is matched.

In order to situate our work within the broader conversation about context in human evaluation of machine translation, Section 2 provides an overview of related work. Since we focus specifically on the case of WMT evaluation as our data source, Section 3 provides the reader with an overview of the changing data collection protocols from the News (later changed to General MT) shared task at WMT from 2018 to 2022. We focus particularly on WMT 2022, which we use for the majority of our analysis, and pay special attention to the calibration set collection at that iteration of the shared task. In Section 4, we begin by taking a closer look at annotator variation using the calibration sets. We show that annotators vary greatly in how they use the space of scores available to them, even when controlling for underlying translation quality, which highlights the importance of calibration sets (especially for in-context evaluation) as well as the need for additional study of annotation protocols. We then focus on the influence of target context on annotator scores in Section 5. Using quality control data, we find that repeat annotations that

occur *after* the annotator has seen degraded target-side intersentential context tend to be assigned slightly lower scores than their counterparts in the original context. We conclude in Section 6 by discussing ways in which annotator behavior, data collection protocols, and user interface design decisions are all intertwined in the results we observe in annotation of machine translation. We provide some suggestions for how to balance some of the conflicting needs in the evaluation of machine translation in context.

2. Background and related work

In 2018, claims of human parity in MT quality (Hassan et al., 2018) saw responses in Toral et al. (2018) and Läubli et al. (2018) that this apparent parity was not replicated when context or other factors like translation direction or annotator expertise are taken into account. Directly inspired by those two papers, WMT 2019 began incorporating source-side document intersentential context into human evaluation of machine translation for its News translation task (Barrault et al., 2019).

Many machine translation systems perform translation at the level of the single sentence, though there is increasing work on context-aware translation (Wang et al., 2017; Jean et al., 2017; Tiedemann and Scherrer, 2017; Voita et al., 2018; Junczys-Dowmunt, 2019; Fernandes et al., 2021, i.a.); see Maruf et al. (2021) or Rauf and Yvon (2022) for a survey of document-level approaches. This is in contrast to the way that human translators work, typically with document context (and world knowledge) available to them. There are some source sentences that can be unambiguously translated in isolation, and whose translation quality can likewise be unambiguously judged even in isolation. However, for many sentences, additional context is necessary for both accurate translation and evaluation. In work on translation quality estimation, Scarton et al. (2015) similarly showed that some issues in human post-editing can only be solved with intersentential context (at the paragraph level) but not when sentences are presented in isolation.

Voita, Sennrich, and Titov (2019) examine this very problem, finding translations that are "good" in isolation but which are revealed to be incorrect when evaluated in context. They perform a two-step manual evaluation process, first annotating sentence translations in isolation and marking them as "good" if they are both fluent and potentially reasonable translations of the source sentence. In the next step, they consider pairs of consecutive sentences and judge their quality again in this paired context. In 7% of these pairs (of English–Russian subtitle data), the individual sentences had been judged to be "good" in isolation but were determined to be bad when viewed together as a pair. This can be viewed as a lower bound on this type of error for this dataset; it remains possible for a pair of sentences to be consistent with one another but nonoptimal in an even broader context.

Three main sources of these context-related challenges, as highlighted by Voita et al. (2019), are deixis (referential expressions whose denotation depends on context), ellipsis (omission of words which are understood via context in the source, but may be necessary or require clarification in the target language), and lexical cohesion (including consistency of terminology). Castilho, Popović, and Way (2020) use an overlapping but slightly different taxonomy across a range of domains and language pairs, noting issues of ambiguity, gender, problems with the source, terminology, and unknown words among the major sources of need for context.

Castilho et al. (2020) and subsequently Castilho (2022) raise the question of how much context span (on the source side) is necessary to accurately translate and evaluate. They find that the number of prior or subsequent source segments needed for accurate translation and evaluation varies based on the type of context needed (whether for gender, ambiguity, etc.) as well as domain and language pair. Castilho et al. (2020) found that many sentences (in some domains a strong majority of sentences) contain portions that cannot be disambiguated by just the preceding two segments. Castilho (2022) found that the median number of context sentences needed tended to be fairly small, though the average could be quite large, indicating that in most cases only a small amount of context is needed, but in some cases the necessary information can only be found

quite far away from the segment of interest. This supports providing as much available context to annotators as can be reasonably balanced against time and annotator fatigue.

In this work, we examine issues of context in evaluation. In particular, we focus on sentence-level evaluation within document context. This category includes the annotation style discussed in Toral et al. (2018)—in contrast to Läubli et al. (2018), who perform comparisons at the full-document level. Given constraints on annotator time, it is hard to collect sufficient document-level scores to reach appropriate levels of statistical power (Graham, Haddow, and Koehn, 2020), while segment-level scores in document context are faster to collect in large quantities. Castilho (2020a) also showed that providing scores at the full-document level is not ideal, due to issues of annotator tiredness and the complexity of giving a single score to whole documents, at least as currently implemented. This preference for segment-level scoring is supported by Castilho (2020b), which compared Likert-style annotator scores at the full-document level to Likert-style annotator scores of randomly ordered single sentences, and found higher levels of inter-annotator agreement on adequacy at the segment level than at the full-document level, with annotators also preferring the former. Expanding this work to examine a three-way comparison of scoring randomly ordered single sentences, scoring individual sentences in context, or scoring full documents, Castilho (2021) found the highest levels of inter-annotator agreement in the randomly ordered single sentence setting, followed by the scoring sentences in context, and the lowest levels when scoring full documents. However, that work also showed that annotators preferred scoring individual sentences in document context. It also speculated that the higher levels of inter-annotator agreement observed in the randomly ordered sentences may not be strictly a good thing: it may be the case that annotators are more permissive of errors or ambiguities (such as ambiguous translations) when they are asked to annotate them without the necessary context. Scoring individual sentences within document context strikes a careful compromise between the benefits of providing annotators with intersentential context and the greater inter-annotator agreement associated with segment-level evaluation. While the annotation data that we will examine in this work is performed using direct assessment (DA; Graham et al., 2013, 2014, 2016) rather than Likert-style scoring, the aforementioned analyses support the approach of segment-level scoring within a document context that is used at WMT.

There are also nuances in how the context is presented to annotators, and the effects those interface choices have on annotation. Toral (2020) raised questions about whether the lack of future document context and the ephemeral nature of prior context produced by the WMT 2019 practice of scoring segments in document order but not allowing annotators to return to the prior segments made the necessary document-level information less available to and less useful to annotators. Grundkiewicz et al. (2021) compare two versions of the user interface used for segment rating in document context at WMT, a "document-centric" version (where all sentences within a document are viewed on the same page and annotators have the option to return to earlier segment scores and update them) and one where sentences are presented in the document order but with only one sentence per page and no ability to return to examine earlier sentences. They find that inter-annotator agreement improves in the "document-centric" approach, though there may also be some evidence of annotator fatigue and increase in annotation time.

3. WMT protocols and data

We focus especially on data from WMT 2022, which included calibration sets scored by multiple annotators, but also examine data from earlier WMT years, dating back to 2018. All scores we examine were collected using variations on DA (Graham et al., 2013, 2014, 2016), a 0–100 sliding scale annotation approach used at WMT since 2016. Here, we describe how those data were collected as a reference for the reader.

The scores are collected in sets of 100–200 segments at a time, often referred to as "HITs" (using Amazon Mechanical Turk terminology for "Human Intelligence Task"). The segment-level scores

are used to compute system-level rankings. First, sometimes at the HIT level and sometimes at the annotator level, the mean and standard deviation of segment-level score are computed. These are then used to compute z-scores, such that for every score x within the given HIT or annotator's data, the new standardized score is $z = \frac{x-\mu}{\sigma}$ where μ and σ are the mean and standard deviation, respectively. After this, for each system, scores of repeated segments are averaged (within and across annotators), and then all segment-level scores for that system are averaged into raw and z-score system averages. For additional details, such as whether quality control items were used in producing the means and standard deviations, see the various system description papers and released code.

3.1. WMT 2018

At WMT 2018 (Bojar et al. 2018), DA scores were collected via Amazon Mechanical Turk (henceforth MTurk) and Appraise (Federmann, 2012), with a mix of crowd and MT researchers as the annotators. Both interfaces had sliding scales ranging from 0 to 100, with tick marks at 25, 50, and 75. The MTurk interface showed the extremes labeled as 0% and 100%, and annotators were asked to indicate to what extent they agreed with the statement "The black text adequately expresses the meaning of the gray text in English.", while the Appraise interface instructions read "How accurately does the above candidate text convey the original semantics of the reference text? Slider ranges from Not at all (left) to Perfectly (right)." Annotators judged one sentence per screen, presented in random order (no document context), in HITs of 100 sentences. Because the sentences were scored in isolation, when multiple systems produced the same output, a single score for that output was then distributed to all systems who had produced that identical output, thus saving annotator effort. HITs are implemented somewhat differently between MTurk and Appraise, so in some settings the smallest unit of collection for Appraise is actually two HITs, or about 200 segments. This can vary between years or language pairs; see the data or system descriptions for more details.

There were three types of quality control: repeat pairs (where annotators were expected to give similar scores), degraded output ("bad reference pairs" where the score was expected to be worse than the corresponding original), and human-translated references (expected to be scored higher than MT output). These "bad reference pairs" are constructed by taking a translation and substituting a sequence of n-grams in it with a sequence of the same length randomly selected from the full test set of reference translations. The expectation is that this substitution will almost always degrade the quality, but in a manner that requires the annotator to read the full sentence to be sure (as the subsequences are expected to be fluent). The standard HIT structure aimed to contain equal proportions of translations from all systems submitted for that language pair. There was also a "non-standard" HIT structure used for some language pairs, which only included the degraded output quality control rather than all three types. All language pairs (except a test pair of English to Czech) were evaluated monolingually, i.e., by comparing the MT output to a reference translation.

3.2. WMT 2019

WMT 2019 introduced document context into the evaluation and performed monolingual (reference-based) evaluation for some language pairs and bilingual (source-based) evaluation for others (Barrault et al., 2019). Annotation was once again performed using MTurk and Appraise. There were two types of document context evaluation: segment rating with document context (SR+DC) and document rating (DR+DC). These were performed sequentially: an annotator would score each segment in a document (all translated by the same MT system) in order and then be presented with a screen showing the full document to provide the document-level score.

The screenshots in the shared task paper show the same sliders and tick marks as in 2018. The style of annotation performed in prior years was renamed segment rating without document context (SR−DC). In this work, we focus only on the segment rating (with or without document context) due to the support for that approach discussed in Section 2, omitting the document rating.

The HITs for SR+DC were constructed by pooling all document–system pairs, sampling from this pool at random without replacement until the HIT contained 70 segments, filling the remainder with quality control items and then shuffling. Note that unlike in SR−DC, there is no attempt made to get coverage of all systems in a HIT (a task that would frequently be impossible). The test datasets from 2019 (with some exceptions) use translations that were produced in the same direction as the machine translation (i.e., translated from the source to the target).

The quality control for the SR−DC data collection remained the same as 2018. The quality control and HIT construction for SR+DC in 2019 are somewhat unclear: it is stated that the same three types of quality control are used for SR+DC; however, those were not contained in the released data and for bilingual evaluation, the reference text is actually treated as another system.

3.3. WMT 2020

At WMT 2020, there was a separation between how annotation is performed for translation into English and translation out-of-English (Barrault et al., 2020). The into-English translations were performed in MTurk, using monolingual evaluation and either SR−DC or SR+DC evaluation, with all three types of quality control items. For out-of-English, bilingual evaluation was performed in Appraise by a mix of researchers and external translators, using SR−DC when document boundaries were unavailable and SR+DC otherwise, and only "bad references" were used for quality control. There was a difference in the user interface for SR+DC in Appraise, where rather than showing each new segment on a separate screen in document order, the annotators had access to the full document context as they scored each segment. They were also able to return to and update segment scores within a document. This addresses the concern described in Toral (2020) about the ephemerality of the earlier document context. In later years, this version of data collection was called SR+FD (segment rating with full document context). The SR−DC and SR+DC/FD HIT sampling procedures remained the same as in the previous year.

3.4. WMT 2021

In 2021, into-English evaluation was performed monolingually on MTurk using SR+DC, while out-of-English and without-English were performed bilingually in Appraise using SR+FD, with HIT construction and quality control the same as 2020 (Akhbardeh et al., 2021). Annotators for the into-English reference-based tasks were crowd workers on MTurk. The annotators for the source-based out-of-English and without-English tasks were from the crowdsourcing platform Toloka, as well as researchers and paid professional annotators.

3.5. WMT 2022

At WMT 2022, data for into-English evaluation was again collected monolingually with MTurk using SR+DC, while there were more changes to the out-of-English/non-English data collection (Kocmi et al. 2022). We focus on the 2022 out-of-English data collection in this work (along with the Chinese–English data that were collected in the same manner).

The first change in 2022 is the introduction of DA combined with scalar quality metric (SQM), which the organizers call DA + SQM. This approach was inspired by Freitag et al. (2021), which found that discrete 0–6 SQM correlated well with multidimensional quality metrics (MQM). In this DA + SQM setup, annotators provided scores on the standard 0–100 sliding scale, but the

scale was marked with 7 tick marks. At the 0, 2, 4, and 6 tick marks, there was text describing the quality level: "0: Nonsense/No meaning preserved", "2: Some meaning preserved", "4: Most meaning preserved and few grammar mistakes", and "6: Perfect meaning and grammar." The instructions on the screen also provided more detailed explanations of each of these quality levels.

It is worth briefly comparing this approach to the adequacy and fluency judgments performed at early iterations of the shared task on machine translation. Koehn and Monz (2006) performed adequacy and fluency judgments with 1–5 scales (for adequacy judgments this ranged from "None" to "All Meaning" and for fluency from "Incomprehensible" to "Flawless English"—these definitions were drawn from a 2005 update to the annotation specifications of LDC (2002)). Noting that different annotators may use the scales differently, they performed normalization. The next year, Callison-Burch et al. (2007) used the same scales and noted that the lack of guidelines given to annotators meant that some annotators used rules of thumb while others treated the scales as relative rather than absolute; their analysis showed lower inter-annotator agreement than relative ranking. Due to this lower inter-annotator agreement as well as the time required for the judgments, Callison-Burch et al. (2008) abandoned the fluency and adequacy scores in favor of relative ranking the following year. While the current SQM and DA + SQM use slightly longer descriptions, it is possible that there may be similar issues in inter-annotator agreement to what was observed in Koehn and Monz (2006) and Callison-Burch et al. (2007).

The next change was in data sampling. Rather than sampling full documents, document "snippets" of consecutive sentences were sampled (approximately 10 consecutive segments, though shorter documents could also be sampled as a whole). Up to 10 preceding and following source segments were also provided in the interface as source context. There was also an attempt made to ensure that all systems were annotated over the same set of segments, a change from prior years (which had left that to chance). For all the language pairs that we focus on in this work, all systems received annotations over all sampled snippets.

3.6. Calibration HITs at WMT 2022

In addition to the other changes described above, "calibration HITs" were collected for six language pairs at WMT 2022: English (eng) into {Czech (ces), German (deu), Croatian (hrv), Japanese (jpn), Chinese (zho)}, and Chinese–English. For each language pair, the calibration HIT consisted of a set of up to 100 randomly selected segments to be annotated (in context) by each annotator in addition to any other annotations that they performed. These annotations were performed in the same DA + SQM setup described in Section 3.5 but did not contain any quality assurance items.

Kocmi et al. (2022) described the calibration HIT data at a very high level, providing a brief summary of correlations between annotators and qualitative commentary on score distributions, but leaving additional analysis to future work; here, we delve into the details. Table 1 shows some summary statistics about the calibration HITs. Each segment in the HIT is associated with a system that translated it, a snippet identifier (composed of the original document identifier and the range of segments in the snippet), as well as the segment's index into the snippet. The count of unique system–snippet–segment triplets is the size of the set of such unique triplets; note that this may differ from the set of unique system–segment pairs, as it is possible for overlapping snippets to have been sampled from the same document (e.g., for one snippet to sample the first 10 segments of the document and another snippet to sample 10 segments starting from a segment included in the first snippet). Depending on the language pair, they range in how many unique system–snippet–segment triplets they cover (93–100; each HIT does contain 100 total segments, the remainder here being repeated from another snippet in the HIT),[b] how many unique document snippets they cover (8–15), how many unique systems they cover (6–12), and the range of segments annotated for the various document

[b]Unless otherwise noted, we average together the scores of duplicates in the various plots shown.

Table 1. Basic calibration HIT statistics

Lang. pair	Unique segs.	# Documents	# Systems	Minimum segs.	Median segs.	Maximum segs.
eng-ces	93	8	6	10	10.0	11
eng-deu	93	9	6	10	10.0	11
eng-hrv	100	13	7	4	7.0	11
eng-jpn	95	9	7	10	11.0	11
eng-zho	100	14	8	2	5.0	11
zho-eng	100	15	12	1	5.0	11

Contents of calibration HITs by language pair: total unique system–snippet–segment triplets in the HIT, number of unique document snippets, number of unique systems, and minimum, median, and maximum number of segments per document–system pair.

Table 2. Basic annotator statistics

		Additional HITs			Additional segments		
Lang. Pair	# Annotators	Minimum	Median	Maximum	Minimum	Median	Maximum
eng-ces	16	1	4.5	29	200	900.0	5800
eng-deu	14	1	6.0	17	200	1200.0	3400
eng-hrv	13	2	6.0	26	388	1200.0	5154
eng-jpn	17	1	4.0	33	200	800.0	6600
eng-zho	8	2	18.0	32	400	3600.0	6400
zho-eng	12	2	7.0	26	400	1400.0	5200

Basic statistics about annotators who completed calibration HITs and additional HITs. For each language pair, the table shows the number of such annotators and the minimum/median/maximum number of additional HITs and additional segments (including quality control segments) that they annotated.

snippet–system pairs covered. The calibration sets, due to their fixed maximum size, do not have full coverage of either documents from the test set or of systems submitted for the language pair. In three language pairs, the calibration set contains repeated document snippets translated by two different systems. For English–Czech, the snippet `ecommerce_en_44#1-10` appears with translations by both HUMAN-B and Online-W, for English–Chinese, `en_zh-TW_CLIENT-05_2020- 12-20-24_doc0#1-5` appears for both Online-B and Online-Y and `en_fr_CLIENT-02_default_2020-12-27-46_doc0#1-5` appears for both Online-W and Lan-Bridge, and for Chinese–English, `en_zh-TW_CLIENT-05_2020-12-20-24_doc0#1-5` appears for both JDExploreAcademy and Online-W.

Table 2 shows information about the number of annotators for each of the language pairs. We omit annotations from annotators who did not complete the full calibration HIT (one each in all language pairs except English into {German, Croatian}) and annotators who did not complete any additional HITs beyond the calibration HIT. For most language pairs, three or fewer annotators completed the calibration HITs without completing any other HITs, but for English–Czech there were 10 such cases. English–Chinese stands out for having a much higher number of median additional HITs per annotator (18) as compared to the other language pairs. The number of additional HITs and segments annotated enables us to reason about the potential costs of using calibration HITs more broadly at WMT. For language pairs like English–Chinese, where the median number of additional HITs completed was 18, requiring annotators to complete calibration HITs represents a relatively small cost if 8 annotators complete them and then most go on to complete a

large number of additional HITs. On the other hand, for a language pair like English–Japanese, 17 annotators complete the calibration HIT and then complete a median number of 4 additional HITs; this means that a larger proportion of the data collected is concentrated in the calibration HIT, rather than spread across the dataset. This trade-off is necessary to keep in mind when considering whether it will be cost-effective to complete a calibration HIT. Future work may also consider how small a calibration set can be in order to still be useful, while minimizing the cost of annotation.

4. Calibration HITs and annotator variation

The main relationship between calibration HITs and context is that collecting calibration data is one potential way to navigate some of the challenges that are introduced to the annotation task by the addition of document context. In order to better understand why we may want to use some form of annotator calibration—particularly in a setting where annotation is performed with intersentential document context—we should first develop a better understanding of annotator variation. We look at the calibration sets collected at WMT 2022 in order to do so. In Section 4.1, we begin by looking at how annotators use the space of possible scores in different ways, before moving on to how annotator scores correlate with one another in Section 4.2. We then discuss how calibration HITs can be used for score normalization in Section 4.3. Finally, in Section 4.4, we examine the implications of what we have learned about annotator variation for the design of machine translation evaluation protocols, particularly those that include document context.

Calibration HITs allow us to examine annotator score distributions in a controlled way. There has been evidence since at least WMT 2006 that annotators vary in the shape of their score distributions (Koehn and Monz, 2006; Koehn, 2009, p. 220). However, limited test set overlap between annotators has made it hard to examine this in detail, disentangled from differences caused by system quality or segment translation difficulty. By examining the calibration HITs, with all annotators scoring the same segments in the same context and order, we are able to more closely examine how annotator score distributions vary while controlling for underlying translation quality, as well as examining correlations between annotators' scores. In the interest of being concise, we will use one language pair (English–Japanese) as the primary example in figures and analyses in the following sections. In most cases, the trends we observe with this language pair are consistent across the other language pairs for which calibration HITs were collected, and we make note of it when they are not.

4.1. Annotator score distributions

Figure 1 shows a subset of the English–Japanese calibration HITs.[c] We provide full versions of this figure and the corresponding figures for other language pairs in the supplementary material. We first focus along the diagonal, where we see a histogram for each annotator, showing the counts of scores that annotator assigned. We look at three interconnected ways that the distributions of scores differ, recalling that all annotators here were annotating *the same set of sentences* in the same context and order. These are score range, distribution shape, and discretization.

All annotators were presented with the same sliding scale from 0 to 100 (with SQM-style tick marks at seven equidistant points on the scale) but not all of them use the full range. For example, annotator `engjpn1611` never gave a score below 52 to any segment, while most other annotators gave at least some scores below 25. This is exactly the sort of issue that z-score standardization is intended to resolve: `engjpn1611` is a *lenient* annotator by comparison to other annotators and has a smaller standard deviation than others. If their scores were well correlated with other annotators' scores (a topic which we discuss below), mapping to z-scores could resolve these inconsistencies.

[c]Figures produced using Matplotlib's Pyplot version 3.6.2 (Hunter, 2007).

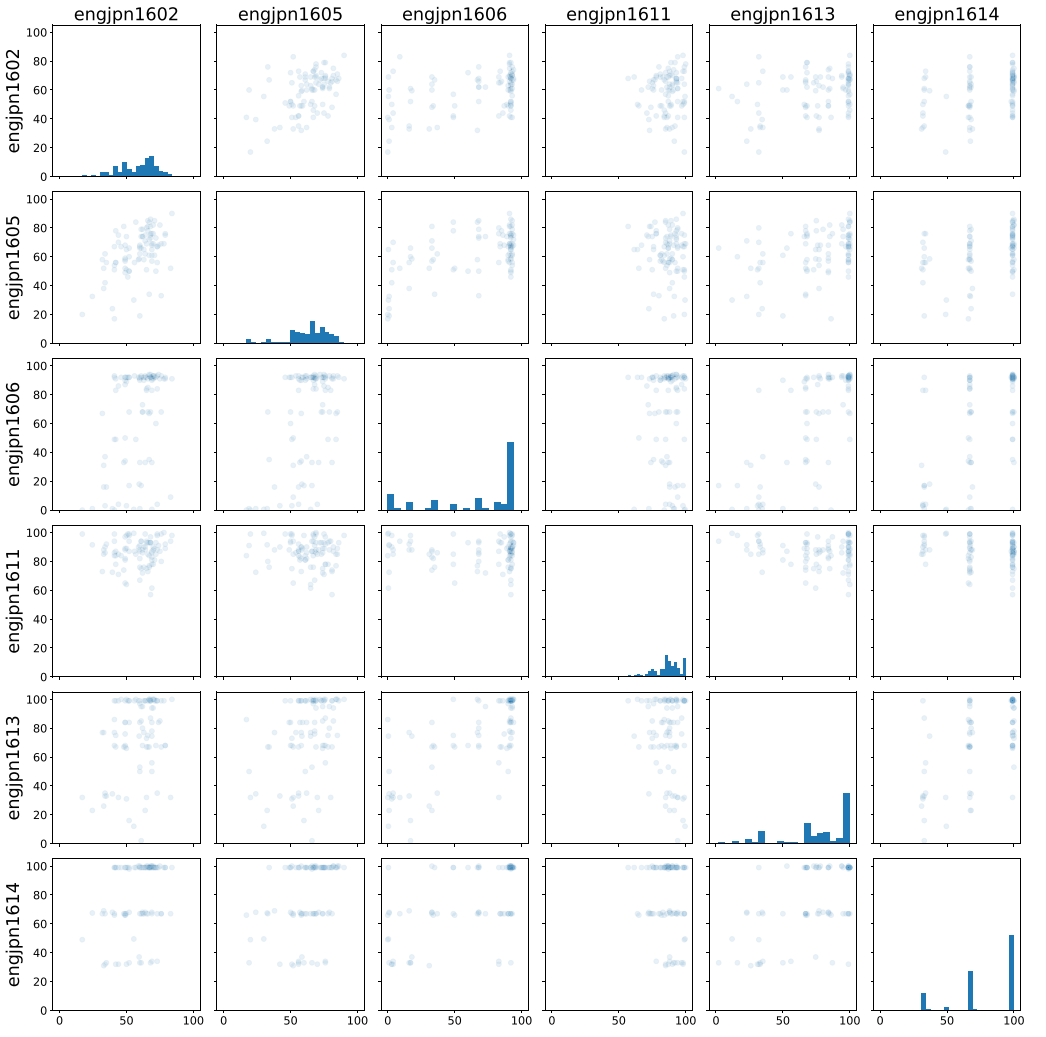

Figure 1. Variations in annotator score distributions and how annotators use the 0–100 scale, shown for a subset of the English–Japanese calibration HITs. Figures along the diagonal show histograms of annotator scores, while the off-diagonal figures show a comparison between two annotators' scores, with each point representing a segment (its x-value determined by the score given to it by the annotator for that column and its y-value determined by the score given to it by the annotator for that row).

A problem that cannot be resolved with z-scores is that of the shape of annotator score distributions: z-score computation preserves the shape, only shifting the mean and scaling to a standard deviation of 1. We see several annotators here whose scores may be considered to be (roughly) normal or skewed normal distributions. In other cases, annotators tend to give most of their scores at the extremes (like annotator `engjpn1606`) and fewer in the middle, or to give few low scores and many high scores (like `engjpn1613`) while still using most or all of the 0–100 scale.

Finally, while DA is often described as being a continuous scale from 0 to 100, in practice it is implemented using discrete integer values. While many annotators make use of the wide range of values available to them as scores, some annotators instead concentrate their scores around a smaller subset of the scores available to them. For example, as shown in Figure 1, annotator `engjpn1614` clustered their scores quite tightly: 53 segments scored between 98 and 100, 31 scored between 66 and 69, 14 between 31 and 34, and 1 each at 41 and 49, with no segments

having any other scores. Put another way, they've assigned just 13 unique scores to the 95 unique segments (100 scores in all) in the calibration HIT, while other annotators assigned upward of 40 unique scores to those same segments. In one even more extreme case (not shown in Figure 1), an annotator gave 2 segments a score of 83, 18 a score of 84, 51 a score of 85, 27 a score of 86, and 2 a score of 87, thus assigning only 5 unique scores, all clustered immediately around the score of 85. It is worth noting how small this particular range is; while there may be differences in implementations and annotator screens, earlier work on DA notes that a 1-point difference may be just 6 pixels on a screen (Graham et al. 2013) and the WMT Metrics task has used a 25-point difference cutoff when performing relative ranking comparisons between DA scores of segments (Mathur et al., 2020).

When annotators choose to discretize the score space by giving scores that are clustered around a small number of points on the scale rather than distributed across the full range, these do not tend to be arbitrary discrete values. Instead, they correspond to features of the user interface. In Figure 1, the annotators who discretized the space do so at approximately the tick marks corresponding to the 0–6 SQM values, or those tickmarks and the midpoints between them. In earlier years, rather than ticks at 0, 16.7, 33.3, 50, 66.7, 83.3, and 100, there were tick marks at 0, 25, 50, 75, and 100. In those years, annotators who discretized the space tended to do so in line with those five tick marks instead.

The small number of repeat annotations in some of the calibration sets offer us a limited view of intra-annotator consistency. As is the case with most other aspects of annotation, intra-annotator consistency varies widely. Most annotators have a median absolute difference of 5 or lower between the two scores of repeated items, while some have median differences as high as 29. For annotators who discretize the space, even if their median absolute difference of repeat segment scores is quite low, their maximum may be high: rather than a slight adjustment, they may move them one or two discrete categories away (e.g., score differences of 16.6 to 33.3). However, large score differences are not exclusive to annotators who discretize the score space; other annotators who distributed their scores in approximately normal distributions sometimes also score the repeated segments quite differently (e.g., score differences of about 20). Of course, with just 5–7 repeated segments to examine in the calibration, it would be unwise to draw extensive conclusions about annotator consistency from this. We perform a more in-depth examination of annotator consistency and the effects of context in Section 5.

The calibration set is quite small, which raises the question of whether observations about an annotator's use of the 0–100 scale will generalize across additional HITs completed. We find, based on a visual inspection across language pairs, that annotators tend to be fairly consistent in the three areas we examined above. They tend to use approximately the same range of scores, maintain similar distribution shapes, and remain consistent in terms of discretization of the space (or lack thereof). Figure 2 demonstrates this for the same set of annotators shown in Figure 1, comparing the distribution of their scores on the calibration HIT and the additional HITs they completed. For this comparison, we include only the scores of real machine translation output, omitting the "bad references" since the calibration HIT contained none of those. As we discuss in Section 5.2, the scores assigned to "bad references" tend to be extremely low, an extreme of quality that was not captured in the calibration set. Thus, we could also imagine scenarios where other extremes of quality not captured in the calibration set might result in changes to the score distribution (e.g., more low scores or more high scores), but it seems to be the case that (with the exception of the "bad references", which most annotators give extremely low scores to) annotators are consistent in whether they discretize the space or not, as well as the general shape and ranges of their score distributions. Nevertheless, this observed consistency is evidence that it is appropriate to treat calibration set annotations as indicative of how broader annotations will be scored—thus calibration set data could be used to select which annotators should or should not complete future annotations, or to indicate whether additional instruction needs to be provided to annotators.

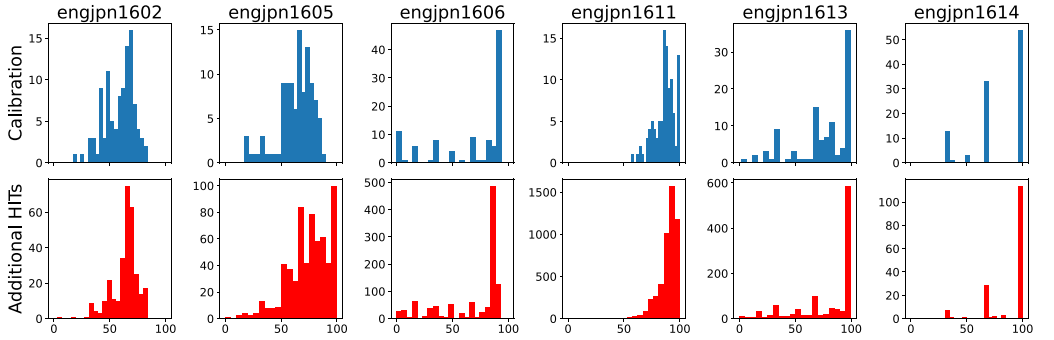

Figure 2. Annotator score histograms across calibration HITs and additional HITs completed ("bad references" omitted), for English–Japanese (same subset of annotators as shown in Figure 1). The x-axis is the score and the y-axis is the count of segments that were assigned that score.

4.2. Score correlations and annotator agreement

The WMT Findings papers sometimes conflate quality control with measuring annotator agreement, arguing that the 0–100 scale prevents the use of standard measures of agreement like Cohen's kappa (Bojar et al., 2018).[d] However, given a calibration set (or other source of annotation overlap), there are still ways that we can examine annotator agreement and inter-annotator consistency beyond simply observing whether they pass quality control. We begin by looking at correlations between annotators' scores.

We examine Spearman's ρ, Pearson's r, and Kendall's τ_c, though for Pearson's r we note that the assumptions of a linear relationship and normal distributions are clearly not met for all annotator pairs. The subset of annotators shown in Figure 1 was specifically selected to include the pairs of annotators with both the highest and lowest Spearman's ρ correlations over their $n = 95$ paired segment scores from the calibration HIT.[e] The annotator pair with the highest correlation was `engjpn1606` and `engjpn1613`, with a Spearman's ρ of 0.69 ($p < 0.001$) and a Pearson's r of 0.75 ($p < 0.001$); though even with this high correlation, there are some points of extreme disagreement), and a Kendall's τ of 0.50 ($p < 0.001$), while the pair of `engjpn1611` and `engjpn1614` had the lowest Spearman's ρ correlation of -0.10 ($p = 0.35$) and the pair of `engjpn1605` and `engjpn1611` had the lowest Kendall's τ of -0.06 ($p = 0.52$). The pair with the lowest Pearson's r was `engjpn1611` and `engjpn1612` (not pictured) with -0.21 ($p = 0.04$). Over the full set of pairs of annotators who completed calibration hits for English–Japanese, the median Spearman's ρ was 0.25, the median Pearson's r was 0.27, and the median Kendall's τ was 0.17.

Table 3 summarizes the correlations between annotators for the full set of calibration language pairs. We see that for some language pairs, there exist pairs of annotators with negative correlations (though these have very high p-values, indicating that there simply may be no correlation, rather than that there is actually a true negative correlation between the annotators' scores). In the case of Japanese, the pairs with the lowest correlations either used less of the available score space (giving only scores above 50) or discretized the score space, highlighting the challenge of comparing scores across this range of possible uses of the score space. Having a set of annotators whose scores are correlated with one another may help ease concerns about the effects of having

[d]While recent WMT Findings papers do discuss correlation between system rankings using different styles of annotation (Bojar et al., 2016, 2018; Barrault et al., 2019) and those produced by crowd workers or researchers (Bojar et al., 2016, 2017), they do not examine correlations between annotators.

[e]Values of Spearman's ρ, Pearson's r, and Kendall's τ were calculated using `scipy.stats.spearmanr`, `scipy.stats.pearsonr`, and `scipy.stats.kendalltau(variant='c')`, to account for ties and the fact that different annotators used different proportions of the available scores), respectively, using SciPy version 1.10.0 (Virtanen et al., 2020).

Table 3. Pairwise annotator correlations

	Spearman's ρ			Pearson's r			Kendall's τ		
Langs.	Minimum	Med.	Max.	Minimum	Med.	Max.	Minimum	Med.	Max.
eng-ces	0.12 ($p = 0.25$)	0.48	0.80	0.21 (0.04)	0.53	0.81	0.08 (0.26)	0.34	0.61
eng-deu	−0.19 ($p = 0.06$)	0.19	0.51	−0.11 (0.31)	0.35	0.77	−0.10 (0.08)	0.11	0.35
eng-hrv	0.20 ($p = 0.04$)	0.40	0.63	0.21 (0.03)	0.55	0.77	0.14 (0.05)	0.26	0.44
eng-jpn	−0.10 ($p = 0.35$)	0.25	0.69	−0.21 (0.04)	0.27	0.75	−0.06 (0.37)	0.17	0.50
eng-zho	−0.10 ($p = 0.34$)	0.14	0.58	−0.12 (0.22)	0.15	0.52	−0.07 (0.33)	0.10	0.42
zho-eng	0.03 ($p = 0.77$)	0.40	0.75	−0.27 (0.01)	0.38	0.73	0.03 (0.72)	0.29	0.57

Minimum, median, and maximum Spearman's ρ, Pearson's r, and Kendall's τ correlations for pairs of annotators, measured over the calibration set. P-values are shown parenthetically. No p-values are shown for medians; all p-values for maximum correlations were $p < 0.001$. (See Table 1 for calibration set sizes.)

different annotators annotate different subsets of the data. However, selecting just for annotator correlation may risk missing out on valid insights and sources of variation from different groups of annotators. Additionally, although a high correlation is not a guarantee that annotators are using the same scale or measuring exactly the same things (e.g., in the case of the many ties that occur for annotators who discretize the space), negative correlations or near-zero correlations are an indication that there is some lack of agreement about either what is being measured or how to score it.

4.3. Using calibration HITs for normalization

While we have established that not all annotator differences can be resolved by z-score normalization, we may still wish to use score standardization, for example, in the case of annotators who have similarly shaped score distributions but quite different ranges (minimum and maximum) of scores.

The data from the calibration HITs provide a natural choice for computing means and standard deviations used for z-scores. The main weakness of such an approach would be that it represents a relatively small number of segments, but the current approach of computing z-scores at the single-HIT level for some language pairs shares this weakness. An alternative is to compute z-scores at the annotator level rather than the HIT level. This is, however, still subject to potential problems if the annotators do not annotate similar distributions of high-quality and low-quality data; the annotator-level means and standard deviations may be influenced by the data itself, rather than solely by annotator behavior. In using the calibration set, we remove this potential source of error: the means and standard deviations for all annotators are computed over a fixed set of data, so the differences that emerge can be understood as differences between annotators rather than being produced by a mix of annotator behavior and distribution of underlying system quality.

In practice, for annotators who complete a large number of annotation sessions, we find that the means as computed over all of their segments tended to be similar to their means over just the calibration HIT, suggesting that the calibration HITs for these language pairs happened to be relatively representative.[f] However, if we take a closer look at each individual HIT completed, we see that there can be quite a bit of variance between their means. For annotators who complete many HITs, it is likely that a reasonable estimate of their mean score over the full test set will be

[f]Note, when we compute these, we do so only including the segments labeled "TGT", omitting the "BAD" reference segments.

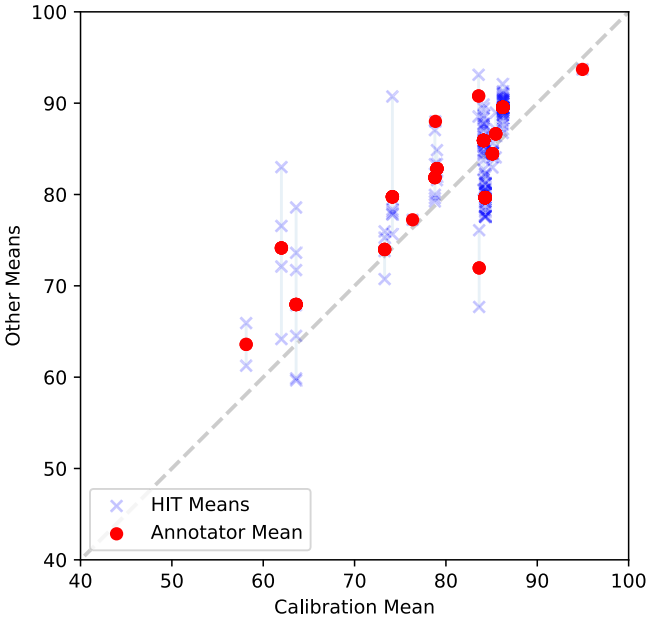

Figure 3. A comparison of calibration means (x-axis), and two groups of other means (y-axis): annotator-level means (dots) and HIT-level means (x marks) for English–Japanese.

produced. For annotators that compute just one HIT, though, there is a risk that their mean is not representative and is instead influenced by that particular HIT (e.g., if it contains no human translations, or happens to contain very low-quality data). As we see in Figure 3, some annotators had annotator-level means that were quite close to their calibration means (i.e., close to the $y = x$ dotted line), while others varied from it, and the individual HITs varied even more. From this data, we cannot determine with confidence how much of those differences are related to intra-annotator (in)consistency as opposed to variation in the underlying quality of the data in the different HITs. While it is possible that the calibration HIT itself is not representative of the full dataset, it at least provides a consistent basis for computing the scores.

To better understand the problem with the current approach of computing HIT-level scores, consider the correlation between an annotator's raw scores and their z-scores. If you use the data from the calibration set to compute μ and σ (mean and standard deviation) and apply it to all of an annotator's scores, their raw scores and their z-scores will have a perfect linear correlation (Pearson's r equal to 1). The same is true if you use all of the annotator's data to compute μ and σ. In both of these situations, you are simply performing a translation (subtracting μ) and scaling by a constant (dividing by σ), so the perfect linear relationship between the raw scores and the z-scores is maintained. However, if you perform z-score normalization at the level of individual HITs,[g] for an annotator who completed n HITs, you will be computing a set of n means and standard deviations: $\{(\mu_1, \sigma_1), (\mu_2, \sigma_2), \ldots, (\mu_n, \sigma_n)\}$. Then for a score x in HIT i, its z-score will be $z = (x - \mu_i)/\sigma_i$, a piecewise function.[h] Now if we imagine that an annotator gave some pair of segments the same score in different HITs, their z-scores may end up being different because they

[g]WMT sometimes computes z-score normalization at the HIT level and sometimes at the annotator level; see Knowles (2021) for discussion of this.

[h]The fact that the application of z-scores is piecewise also contributes to why the raw scores and z-scores at the system level are not perfectly correlated; though in that case it is *intended* to be a feature and not a bug, since its goal is to normalize away annotator differences.

happened to occur in HITs with a higher or lower mean score. In practice, we can take the raw scores for each annotator and compare them against their HIT-level z-scores, and we see that this hypothetical concern does occur: in the extreme cases, the Spearman's ρ between an annotator's raw scores and their HIT-wise normalized z-scores drops as low as 0.76; in most cases, it remains above 0.9 (and for annotators who only completed on HIT, it of course remains 1.0). Either computing μ and σ from the calibration HIT or from an annotator's full set of scores will solve this problem; doing so from the calibration HIT has the added benefit of ensuring that the differences being standardized away are annotator differences rather than system quality differences.

We also note that using z-scores are not the only possible approach to annotator normalization. Instead of computing μ and σ, one could instead choose to use the median and interquartile range, for example. A benefit of this would be the use of a statistic that is more robust to outliers.

The main difficulty of using calibration HITs in this way is that it does introduce the requirement to be able to map from HITs to annotators (via anonymous IDs). We argue that this additional effort is worthwhile in order to produce more trustworthy data. It also has the added cost of asking all annotators to annotate the calibration set, which means there will be a large number of overlapping annotations for this small set; if there is a large number of annotators each completing only a small number of annotations, this cost may be problematic, but it represents a smaller fraction of the cost if a small number of annotators are completing a large number of annotations.

4.4. Implications for machine translation evaluation protocols

Annotators are using the space of possible scores in strikingly different ways. Some of these differences may be able to be smoothed over by normalization approaches, but others cannot be. In practice, what we observe is that some annotators are using a wide range of score values, while others are functionally treating it as a discrete scale with only a small number of possible scores. Put another way: some annotators are doing what we might think of as traditional DA (using a larger portion of the 0–100 scale), while others are effectively using a 5 or 7 item rating scale. These very different score distributions are then combined together and treated as though they were all continuous scores. Typically, when researchers in this area want to compare DA scores and discrete scores, they first perform some sort of discretization on the DA scores or compare them on a downstream task like system ranking or pairwise comparison (Graham et al., 2013; Bojar et al. 2016, i.a.); that is not being done in the current approach to producing system rankings.

When collecting this data, we should consider how best to approach this issue in evaluation with intersentential context. With this shift to document-level annotation, there is less ability to ensure that annotators all cover equal amounts of data from each system, making it more important to try to ensure annotator consistency or similarity to start with (Knowles, 2021). For our current approaches to annotation that incorporate document context, we are faced with a challenge: if each annotator annotates n segments (where n may be on the order of 1000, but possibly as low as 200, as we see in Table 2) and these segments appear in blocks of context, even as short as 10 segments, it may be impossible to have an annotator score examples from all systems at all, let alone in equal proportions. Thus, the earlier assumptions about all systems being equally impacted by each individual annotator may no longer hold true or may not hold true as consistently.[i] We consider three approaches to this issue in this setting.

One way to do this would be to use calibration data to select only annotators who use the space of possible scores in a similar way, or only annotators who reach a certain level of correlation or agreement. A risk of this is that subselecting annotators based on correlation or agreement could

[i]One alternative solution to this is to have an annotator score the same set of segments across all systems; however, this has a high risk of annotator fatigue (seeing the same sentence over and over, especially if systems produce similar outputs), so we do not consider it a feasible solution.

wash away valid inter-annotator differences (such as dialect variations, which might or might not be perspectives we wish to keep in the annotation process, depending on the purpose of the evaluation and the target users of the machine translation system).

The second option would be to discretize all scores (since it is possible to discretize the less-discrete scores, but not possible to undiscretize other than by, e.g., adding random noise, which would add another source of uncertainty). This would still have issues of variation in distribution (as was observed at WMT 2006) and would also suggest that it might be better to simply change the interface. However, simply switching to a discrete scale runs the risk of reintroducing the lack of agreement that caused WMT to abandon it in the early years; it likely needs to be approached with clearer instructions to the annotators.

The third option would be to provide more explicit instruction about the desired use of the score space. Should annotators try to provide discrete scores? Should they try to use more of the scoring space to provide nuance? Currently, this appears to be quite underspecified in instruction: an annotator who sees the sliding scale and assumes they are welcome to use all of it and an annotator who sees the tick marks or quality-level descriptions and assumes they should stay close to those both appear to be making very reasonable assumptions given the available information. Clark et al. (2021) suggests three simple approaches to quick annotator training that could be modified to apply to machine translation: instructions, examples, and comparison.

The instructions approach is closest to what is being done in the current DA+SQM interface, though it could be modified to also address the use of the slider, tick marks, and score space. However, it remains to be seen whether this could overcome annotator preferences. This could be done in a manner that is applicable across language pairs. Using the examples approach, a small set of translations could be selected, and a short description of their errors and an approximate expected range of scores could be provided; in this case, rather than descriptions of the SQM categories as in the current instructions, there could also be examples for each of those categories. In the comparison approach, one could give examples of pairs of sentences and show why they differ in desired scores. These latter two approaches would require different examples for each language pair for the training to be most closely aligned with the specific annotation task. Licht et al. (2022) show an example-based approach in their scoring rubric for evaluating cross-lingual semantic similarity; that work also uses calibration sets, but with an explicit goal of producing scores that will be cross-lingually comparable, something that is beyond the scope of current WMT approaches.

If one were to choose the first option of selecting annotators based on their scores of a calibration set (alone or in conjunction with providing more annotator training), there are a number of things to keep in mind. The shape of a distribution on a calibration set is not dependent solely on the individual annotator, but rather it is (ideally) a function of the quality of the systems whose output is included in that calibration set. For this reason, it may be worthwhile to manually (or at least non-randomly) assemble a calibration set, to ensure that it contains a somewhat representative distribution of scores (e.g., including human reference, high- and low-quality systems, etc., perhaps as estimated by automatic metric scores). If the calibration set consists only of data in the extremes (for example, high-quality human translation and very poor machine translation), we would expect annotators to have scores distributed at the extremes, whereas a calibration set consisting of mainly middle-of-the-road quality MT might be more likely to have annotators produce normal distributions of scores. It also isn't necessarily sufficient to use a single metric like correlation (you could have perfect correlation with one annotator who gives a score of n to all segments and one annotator who gives a score of m to all segments, but this is unlikely to be an informative set of annotations).

There is another potential benefit of using calibration sets, albeit one whose benefits this data does not allow us to measure. According to Bojar et al. (2018), "it is known that judges 'calibrate' the way they use the scale depending on the general observed translation quality." By having all annotators first complete a calibration set, ideally one that covers the range of expected machine

translation quality of the full test set and is somewhat representative of the test set as a whole, all annotators are first exposed to a common set of "observed translation quality", helping them to set their expectations for the task ahead. In this sense, the calibration set can also be seen as calibrating the annotator's understanding of the range of quality that they may expect to observe throughout the annotation process, giving them a common frame of reference to start from. It remains an open question whether annotators will calibrate-as-they-go, shifting their scores if they encounter new distributions of quality; in such a case, it could be necessary to reconsider the appropriateness of using calibration sets for standardization. While this is worth future study, it is beyond the scope of what we can analyze with the available data.

5. Influences of target language context on scores

In the previous sections, we have discussed a number of sources of annotator variation and discussed the need to consider how the introduction of intersentential context (in the form of sampling whole documents or document snippets) interacts with assumptions about annotation behavior. Here, we examine how variations in preceding target-side interesentential context may affect annotator behavior. Unlike in the calibration work, where we were primarily interested in *inter-annotator* variation over a fixed set of data, here we are interested in how changes in preceding target-side intersentential context can result in *intra-annotator* score variation.

We first discuss scores of repeated annotations produced without context in Section 5.1 before introducing the degraded context setting that allows us to ask questions about the influence of target-side intersentential context on annotation scores in Section 5.2. We then examine this at the language pair and individual level in Section 5.3, before concluding with analysis and implications in Section 5.4.

Much of the work on context in human evaluation of machine translation has focused on the source-side context needed for adequate disambiguation (Scarton et al., 2015; Voita et al., 2019; Castilho et al., 2020; Castilho, 2022, i.a.), or on how issues of consistency and discourse allow annotators to distinguish between human and machine translation at the system level (Toral et al., 2018; Läubli et al., 2018, i.a.). However, there has been little work examining how changes to *target* context influence segment-level scores; Läubli et al. (2018) provides a brief discussion of the potential influence of target language lexical coherence but does not specifically design experiments to analyze this more closely. We take advantage of the way quality control items were implemented in the 2022 data we have already examined in order to take a closer look at this issue. We then compare to prior years with both document-level and segment-level annotation. This provides us with greater insight about the ways in which context influences human annotation scores at the segment level and allows us to propose a new best practice for future quality control.

5.1. Scores of repeated annotations

In an ideal measurement scenario, annotators scoring MT output would be perfectly consistent: if they were asked to rate a segment twice, they would give it identical scores (assuming consistent context, etc.). In practice, we know this is not the case. Nevertheless, there is an expectation that annotators are relatively consistent (and unlikely to regularly provide extremely divergent scores to the same segment). Thus, if we examine segment scores paired with repeats, we expect to find the difference between the original segment and the repeat pair to be distributed around a mean of 0. We first examine this in the case of the SR−DC annotations at WMT 2018-2020, pooling all language pairs and years together.[j] In all years, while there are some extreme differences (absolute

[j]Specifically, we use the MTurk data which can be found at https://www.computing.dcu.ie/~ygraham/newstest2018-humaneval.tar.gz, https://www.computing.dcu.ie/~ygraham/newstest2019-humaneval.tar.gz, and https://www.scss.tcd.ie/~ygraham/newstest2020-humaneval.tar.gz, as these contain repeat annotations. We omit the Appraise data (which does not contain segments labeled as repeats).

differences as large as 100), both the median and mode values of the differences are 0, as expected from reasonably consistent annotators. We perform a T-test for the mean of the differences on the data (pooling all language pair and all years) and find that the null hypothesis (that the mean is equal to 0) cannot be rejected;[k] annotators' scores tend to be relatively consistent on the whole.

5.2. Effects of target context degradation on scores of repeated segments

In WMT 2022 DA+SQM ratings in document context, as was also the case in SR+DC/FD, the quality control documents consisted of a mix of target sentences that were left untouched and sentences that were degraded ("bad references"). Thus, some of them are truly "bad references", while others are repeated segments scored in a degraded context. Noting that annotators rarely return to revise scores after giving them (Grundkiewicz et al., 2021), we divide these repeated (untouched) segments into two categories: those that appear before the first quality-degraded segment in the document and those that appear after at least one degraded segment.

We hypothesize that the segments that appeared before any bad references will behave approximately like the "repeat" segments that were observed in isolation, that is, that the difference between them will be distributed around a mean of 0. The segments that appear *after* at least one degraded segment, on the other hand, we expect to be (more strongly) influenced by the presence of degraded context. While we could imagine different annotator reactions to segments in degraded context (e.g., giving a higher-than-usual score to the non-degraded segments because they look better by comparison), we expect that overall, annotators will give lower scores to segments when they are observed in a degraded document context.

We first look at WMT 2022 (focusing on the restricted set of language pairs and annotators who completed calibration HITs), where the effect of context might be expected to be strongest, both because of the easier access to context as argued by Toral (2020) and Grundkiewicz et al. (2021) and because they only use "bad references" (and not reference data) as quality control. While the MTurk data is presented in the original HIT order, the Appraise data is not always. We first identify potential pairs where there was a "bad reference" document-level score. We then check within each HIT containing such potential pairs and check whether the annotations for both the original version and the quality control version appear (either whole or in part), and do so in order within the snippet (we discard the cases where they were not in order). We also identify the point in the quality control version of the snippet where the first "bad reference" appears so that we can categorize the repeated segments based on whether they appear before it or after it. Table 4 shows an example of a quality control document and its original document pair, labeled as described. This is a fairly simple approach; it may also be worth considering whether the segments are directly adjacent or whether more than one degraded segment has appeared.

For each repeat segment, we subtract the score in the degraded context from the score in the original, non-degraded context. A positive value indicates that the segment receives a lower score in the degraded context. Again pooling across all language pairs, we find a mean of −0.26 for the difference in scores of the set *before* any degraded context (2378 score pairs) and a mean of 1.9 for the set where the repeat was observed *after* at least one degraded segment (8924 score pairs). We find using a one-tailed T-test that we can reject the null hypothesis with $p < 0.01$; the mean of the differences in the after-seeing-degraded-segments set is greater than the mean in the before-seeing-degraded-segments set.[l] This means that, after seeing a degraded segment, there is a bias toward giving lower scores to the repeat segments. In both cases, the medians and modes are both 0, but we can still conclude that this context influences at least some of the annotators sufficiently to lower scores of the same segments based on observing them in a degraded context. Figure 4

[k]Computed using `scipy.stats.ttest_1samp`.

[l]Computed using `scipy.stats.ttest_ind` with `equal_var=False` and `alternative='less'` to perform the one-tailed T-test.

Table 4. Example quality control document pair

Original document		Quality control			
Label	Score	Label	Score	Before/After	Notes
TGT	85	TGT	85	BEFORE	Repeat annotation before any “bad reference”
TGT	99	BAD	0	-	First instance of “bad reference”
TGT	33	BAD	16	-	
TGT	84	BAD	33	-	
TGT	100	TGT	67	AFTER	Repeat annotation after first “bad reference”
TGT	100	BAD	16	-	
TGT	85	TGT	84	AFTER	Repeat annotation after at least one “bad reference”
TGT	85	TGT	84	AFTER	Repeat annotation after at least one “bad reference”
TGT	100	TGT	100	AFTER	Repeat annotation after at least one “bad reference”
TGT	92	TGT	93	AFTER	Repeat annotation after at least one “bad reference”
TGT	100	BAD	16	-	

Example of a quality control document pair, showing the mix of degraded “bad reference” (BAD) segments and untouched (TGT) segments. All original document segments are labeled TGT. The TGT quality control segments that occur in the document before any BAD segments are labeled BEFORE (this example only contains one, but there can be a sequence of several), while any that appear after the first instance of a BAD segment are labeled as AFTER. This example comes from English–German annotations (engdeu1613).

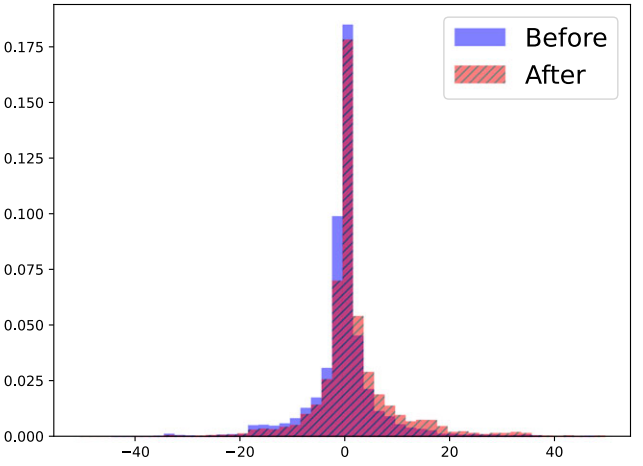

Figure 4. Density histograms showing the shift in score differences between repeat segments observed before encountering any “bad references” (blue/unhatched) and those after encountering “bad references” (red/hatched) for the 2022 data. The horizontal axis shows the score difference (original segment score minus the score in the quality control portion), while the vertical axis shows the fraction of pairs with that score difference. In both the before and after cases, we still observe that the most common difference is 0, but the rightward shift and thicker right tail in the red/hatched after-seeing-degraded-segments set indicates an overall bias toward scoring repeat segments lower in a degraded context.

shows this with overlapping histograms: blue (unhatched) for the difference in before-seeing-degraded-segments pairs and red (hatched) for the difference in after-seeing-degraded-segments. The rightward shift and thicker right tail of the red (hatched) histogram shows the slight general tendency to give lower scores in degraded context which we observed in the 2022 data.

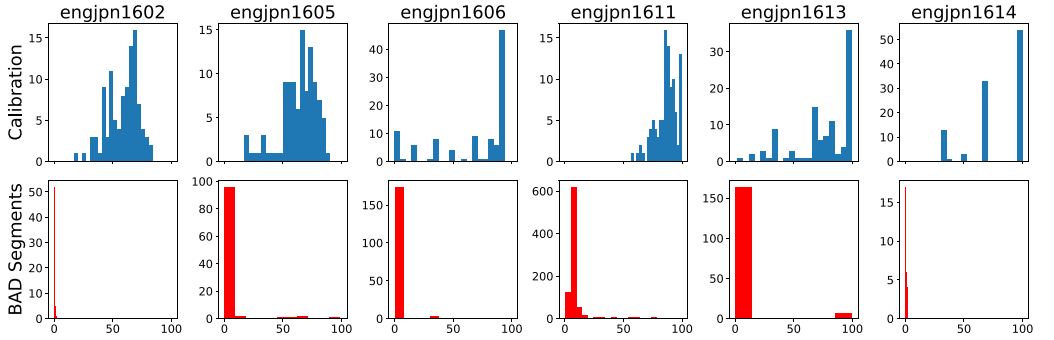

Figure 5. Paired histograms showing the calibration HITs (top) and scores for "BAD reference" items (bottom). This shows the same subset of annotators as Figure 1, with the x-axis again representing the 0-100 score range.

We also examine this for MTurk SR+DC data from WMT 2019 and 2020 data and find the same result as WMT 2022, where repeated segments before any "bad references" behave like the true repeated segments in the SR−DC setting, while those after "bad references" receive lower scores on average. These are again small but significant differences and are somewhat more surprising to observe in this setting because these documents also have reference text (higher quality) mixed in as quality control. Is it simply a matter of inconsistency within the document driving down scores, or are other issues like the ratio of the different types of quality control items playing a role? In WMT 2021 MTurk SR+DC data, we do not find a significant difference between the two distributions (in fact, while the mean of the after-seeing-degraded-segments set is significantly different from 0 and the other is not, the mean of the latter is 0.05 larger than the former). It would require additional study to gain more insight into why we still observed this difference some of the time even in the mixed quality control setting.

We note that the quality of "bad references" is actually fairly extreme: annotators do not score these as similar to other low-quality machine translation output; they concentrate their scores very near 0. This can be seen in Figure 5, where we show the calibration score distributions alongside the score distributions for just the "bad references" in the remaining HITs completed by the selection of annotators. It may be the case that these errors are viewed as so egregious that they poison the perception of the system as a whole or the segments around them. This could be related to the decrease in user trust of machine translation systems after users encounter errors, as described in Martindale and Carpuat (2018).

5.3. Behavior at the language pair and individual level

Focusing on the more clear-cut situation of WMT 2022 DA+SQM, we first break this down by language pairs and then briefly discuss individual annotators. All language pairs except English–Japanese also exhibit the same significant difference between the groups of scores as we saw for the full data. For English–Japanese, both means are small but positive, with the before-seeing-degraded-segments mean actually being slightly larger (0.86) than the after (0.74), though the after is the only one that is statistically significantly different from 0. As we're starting to look at smaller sets of data, and without additional data collection, it's difficult to pinpoint a cause for these differences. It could be the case, for example, that annotators are reading beyond the current segment before giving a score. Such behavior could have an effect on the score (but the timing data does not tell us whether it happened; either a different interface or eye-tracking might be needed to measure this). The strength of the effect of context may also depend on how necessary context is for scoring the data, information that isn't currently collected for the WMT data. For example, it could be the case that annotators pay more attention to context when the data requires

context for disambiguation; without additional study or annotation of the data, this is not possible to determine.

We have too few data points to examine this effectively at the annotator level, but a qualitative analysis suggests that some annotators have minimal differences between pre- and post-"bad reference" repeated segments, while others exhibit the same skew as seen overall. Additional study would be required to examine this in more depth and address questions about why this occurs, such as whether some annotators focus more on context than others.

5.4. Analysis and implications for best practices

This analysis of repeated annotations in degraded context provides additional insight into trends in how target-side variation has an impact on segment-level scores. By holding the source context constant, and degrading parts of the target context, we observe that scores of the non-degraded following sentences also tend to decrease. This could be due to lack of consistency due to the degraded segments, or it could be the case that degraded context predisposes annotators to judge the subsequent sentences more harshly.

There are limitations to the claims we can make from our analysis so far. We are using existing data that has a set of desirable properties for examining this phenomenon, but we are constrained by the limited size and the randomness of the location of the bad references. A more controlled experiment may be able to draw stronger conclusions or to experiment with questions such as the effect of the length of the sequences of degraded context or the effect of different types of magnitudes of degradation.

The main conclusion for best practices here is that—for both theoretical and practical reasons—the scores of repeated segments that are mixed into quality control documents should not be used in calculating overall system scores. The theoretical reasons for this are quite clear. If we believe that document context is important for accurately scoring individual segments, and that document context affects these scores (a claim supported by extensive evidence), then scoring a segment in degraded context is not a fair representation of the system's quality (at the document level or the segment level). Practically, across a range of language pairs, we observe a tendency to give lower scores to these segments in degraded context. If those scores are then used to compute overall system scores, systems that happen to have been used in quality control items risk being penalized more that those that were not.

6. Conclusions

In this work, we have performed a close examination of the WMT 2022 calibration sets, highlighting various topics about annotator variation in this task. We have also examined the effect of target-side degraded intersentential context, finding that it can have an impact on annotator scores. To do this, we have utilized existing datasets from WMT; while the calibration sets were designed to enable this kind of analysis, the quality assurance tasks were repurposed for this analysis. This limits us in the types of generalizations we can make from this data. For example, we cannot make strong claims about whether future data collection should use DA, SQM, MQM, or other approaches. While there has been some comparison of these approaches (Freitag et al., 2021), we encourage additional research into this question.

We do expect, however, that some of the conclusions we draw here are likely to generalize to future evaluations with various different annotation approaches. For example, the work demonstrating the effect of artificially degrading target-side text (for quality control purposes) on the scores of non-degraded, authentic target-side output indicates that researchers should be careful in ensuring that their quality control measures are separated from and unlikely to influence the scores assigned to real output. Additionally, the issues that we highlight about the distribution of tasks to annotators and the need for calibration are likely to arise in a number of settings.

Depending on the level of control over the annotation process, it may also be possible to ensure better distribution of annotators over systems or documents (e.g., via block designs). While this data cannot answer questions about whether giving additional directions to annotators can help produce greater consistency, we would encourage more research into this area, examining how annotators interpret the directions they receive. One form such study could take would be to elicit annotator feedback about how they interpret the task instructions, similar to the approaches used in Shirali-Shahreza and Penn (2023) to collect and cluster different annotator perspectives on the interpretation of "naturalness" in text-to-speech evaluation. Annotator selection, annotator calibration, and instructions to annotators can all be seen as approaches to handling the variation that we observed between annotators; this work provides some considerations for annotation protocol designers, but cannot, with the data provided, answer the question of how best these can be combined, or which is most effective on its own.

We have shown that annotators are quite consistent in some features of their annotation styles (such as the range of scores, whether they discretize the score space or not, and the general shape of their score distributions). This supports the use of calibration sets for normalization and smoothing out inter-annotator differences, though there are still additional considerations (such as the best approaches to normalization, or whether calibration sets should be used as a way to select annotators). We have also demonstrated the effect of (artificially) degraded target-side context on the scores of repeated segments, a finding that provides additional evidence about the influence of context on annotation and which has implications for best practices in quality control.

We conclude by highlighting the main takeaways of this work, in light of the field's necessary shift toward document-level annotation:

- The combination of document length and annotator time/fatigue means that normalization approaches that worked well for segment-level annotation in isolation now risk introducing errors in context-aware annotation (Section 4).
- We should use calibration sets to pre-screen annotators and as a consistent basis for performing normalization (Section 4.3).
- It may be necessary to provide additional training to annotators, for example, via clearer instructions or examples. Critically, this may mean considering the specific goals of annotation, defining standard expectations, and providing more detailed instructions to annotators (Section 4.4).
- Repeat annotations of segments in degraded target-side context (i.e., after "bad references") result in lower scores for those segments, so if degraded target context is going to be used for quality control, those interwoven repeated segments should not be used in computing the overall system scores. This also provides a demonstration of one of the ways target-side context influences annotation (Section 5).

We encourage future work on human annotation in the areas that we did not explore in this paper. A closer analysis of how annotators interpret task instructions and the effects of changing the task instructions on annotator behaviors (e.g., whether more explicit instruction may encourage more annotators to use the full score space, rather than clustering around the tick marks) would help to ensure that we are measuring what we aim to measure and enable us to make changes if we are not. Additional controlled study of different annotation methods and their trade-offs would also benefit the community. To thoroughly answer questions about the effects of differences in score distributions (e.g., shapes of distributions and discretization), an ideal study might collect both small calibration sets and multiple or even all-annotator scores over the remainder of the data to be annotated, which would allow for careful experiments about the effects of different types of standardization approaches or the relative consistency of different groups of annotators. While some of these have been tested with automatic metrics standing in for

annotators, it is not always clear how well those results will generalize. As a research community, we would benefit from continued work toward building and analyzing annotation protocols that enable annotators to produce accurate and consistent judgments, in ways that take into account necessary context while also minimizing annotator fatigue.

Acknowledgments. We thank the reviewers for their exemplary reviews; their thorough and thoughtful comments and suggestions were influential in improving this work. We thank Tom Kocmi and Roman Grundkiewicz for comments and questions, and our colleagues, particularly Michel Simard, Cyril Goutte, and Gabriel Bernier-Colborne, for their feedback.

Competing interests. Rebecca Knowles, the first author of this paper, is a co-editor of the special issue to which this manuscript was submitted; she recused herself from any decisions on the paper's reviewing and acceptance or rejection.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/nlp.2024.5

References

Akhbardeh F., Arkhangorodsky A., Biesialska M., Bojar O., Chatterjee R., Chaudhary V., Costa-jussa M. R., España-Bonet C., Fan A., Federmann C., Freitag M., Graham Y., Grundkiewicz R., Haddow B., Harter L., Heafield K., Homan C., Huck M., Amponsah-Kaakyire K., Kasai J., Khashabi D., Knight K., Kocmi T., Koehn P., Lourie N., Monz C., Morishita M., Nagata M., Nagesh A., Nakazawa T., Negri M., Pal S., Tapo A. A., Turchi M., Vydrin V. and Zampieri M. (2021). Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, pp. 1–88.

Barrault L., Biesialska M., Bojar O., Costa-jussà M. R., Federmann C., Graham Y., Grundkiewicz R., Haddow B., Huck M., Joanis E., Kocmi T., Koehn P., Lo C.-k., Ljubešić N., Monz C., Morishita M., Nagata M., Nakazawa T., Pal S., Post M. and Zampieri M. (2020). Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, pp. 1–55.

Barrault L., Bojar O., Costa-jussà M. R., Federmann C., Fishel M., Graham Y., Haddow B., Huck M., Koehn P., Malmasi S., Monz C., Müller M., Pal S., Post M. and Zampieri M. (2019). Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Florence, Italy: Association for Computational Linguistics, pp. 1–61.

Bojar O., Chatterjee R., Federmann C., Graham Y., Haddow B., Huang S., Huck M., Koehn P., Liu Q., Logacheva V., Monz C., Negri M., Post M., Rubino R., Specia L. and Turchi M. (2017). Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark: Association for Computational Linguistics, pp. 169–214.

Bojar O., Chatterjee R., Federmann C., Graham Y., Haddow B., Huck M., Jimeno Yepes A., Koehn P., Logacheva V., Monz C., Negri M., Névéol A., Neves M., Popel M., Post M., Rubino R., Scarton C., Specia L., Turchi M., Verspoor K. and Zampieri M. (2016). Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Shared Task Papers*, Berlin, Germany: Association for Computational Linguistics, vol **2**, pp. 131–198.

Bojar O., Federmann C., Fishel M., Graham Y., Haddow B., Huck M., Koehn P. and Monz C. (2018). *Findings of the 2018 conference on machine translation (WMT18)*. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Belgium, Brussels: Association for Computational Linguistics, pp. 272–303.

Callison-Burch C., Fordyce C., Koehn P., Monz C. and Schroeder J. (2007). *(Meta-) evaluation of machine translation*. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic: Association for Computational Linguistics, pp. 136–158.

Callison-Burch C., Fordyce C., Koehn P., Monz C. and Schroeder J. (2008). *Further meta-evaluation of machine translation*. In *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, Ohio: Association for Computational Linguistics, pp. 70–106.

Castilho S. (2020a). *Document-level machine translation evaluation project: Methodology, effort and inter-annotator agreement*. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal: European Association for Machine Translation, pp. 455–456.

Castilho S. (2020b). *On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation*. In *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, pp. 1150–1159.

Castilho S. (2021). *Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation*. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. Association for Computational Linguistics, pp. 34–45.

Castilho S. (2022). *How much context span is enough? examining context-related issues for document-level MT*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, pp. 3017–3025.

Castilho S., **Popović M. and Way A.** (2020). *On context span needed for machine translation evaluation*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, pp. 3735–3742.

Clark E., **August T.**, **Serrano S.**, **Haduong N.**, **Gururangan S. and Smith N. A.** (2021). *All that's 'human' is not gold: Evaluating human evaluation of generated text*. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 7282–7296.

Federmann C. (2012). Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics* **98**, pp. 25–35.

Fernandes P., **Yin K.**, **Neubig G. and Martins A. F. T.** (2021). *Measuring and increasing context usage in context-aware machine translation*. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 6467–6478.

Freitag M., **Foster G.**, **Grangier D.**, **Ratnakar V.**, **Tan Q. and Macherey W.** (2021). Experts, errors, and context: a large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics* **9**, pp. 1460–1474.

Graham Y., **Baldwin T.**, **Dowling M.**, **Eskevich M.**, **Lynn T. and Tounsi L.** (2016). *Is all that glitters in machine translation quality estimation really gold?*. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, pp. 3124–3134, The COLING 2016 Organizing Committee.

Graham Y., **Baldwin T.**, **Moffat A. and Zobel J.** (2013). *Continuous measurement scales in human evaluation of machine translation*. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, Sofia, Bulgaria: Association for Computational Linguistics, pp. 33–41.

Graham Y., **Baldwin T.**, **Moffat A. and Zobel J.** (2014). *Is machine translation getting better over time?*. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden: Association for Computational Linguistics, pp. 443–451.

Graham Y., **Haddow B. and Koehn P.** (2020). *Statistical power and translationese in machine translation evaluation*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 72–81.

Grundkiewicz R., **Junczys-Dowmunt M.**, **Federmann C. and Kocmi T.** (2021). *On user interfaces for large-scale document-level human evaluation of machine translation outputs*. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. Association for Computational Linguistics, pp. 97–106.

Hassan H., **Aue A.**, **Chen C.**, **Chowdhary V.**, **Clark J.**, **Federmann C.**, **Huang X.**, **Junczys-Dowmunt M.**, **Lewis W.**, **Li M.**, **Liu S.**, **Liu T.**, **Luo R.**, **Menezes A.**, **Qin T.**, **Seide F.**, **Tan X.**, **Tian F.**, **Wu L.**, **Wu S.**, **Xia Y.**, **Zhang D.**, **Zhang Z. and Zhou M.** (2018). Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Hunter J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering* **9**(3), pp. 90–95.

Jean S., **Lauly S.**, **Firat O. and Cho K.** (2017)). Does neural machine translation benefit from larger context? *arXiv preprint arXiv: 1704.05135*.

Junczys-Dowmunt M. (2019). *Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation*. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Florence, Italy: Association for Computational Linguistics, pp. 225–233.

Knowles R. (2021). *On the stability of system rankings at WMT*. In *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, pp. 464–477.

Kocmi T., **Bawden R.**, **Bojar O.**, **Dvorkovich A.**, **Federmann C.**, **Fishel M.**, **Gowda T.**, **Graham Y.**, **Grundkiewicz R.**, **Haddow B.**, **Knowles R.**, **Koehn P.**, **Monz C.**, **Morishita M.**, **Nagata M.**, **Nakazawa T.**, **Novák M.**, **Popel M. and Popović M.** (2022). Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 1–45.

Koehn P. (2009). *Statistical Machine Translation*. Cambridge University Press.

Koehn P. and Monz C. (2006). *Manual and automatic evaluation of machine translation between European languages*. In *Proceedings on the Workshop on Statistical Machine Translation*, New York City: Association for Computational Linguistics, pp. 102–121.

Läubli S., **Sennrich R. and Volk M.** (2018). *Has machine translation achieved human parity? a case for document-level evaluation*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, pp. 4791–4796.

LDC (2002). Linguistic data annotation specification: Assessment of fluency and adequacy in arabic-english and chinese-english translations.

Licht D., **Gao C.**, **Lam J.**, **Guzman F.**, **Diab M. and Koehn P.** (2022). *Consistent human evaluation of machine translation across language pairs*. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, Orlando, USA, pp. 309–321, Association for Machine Translation in the Americas.

Martindale M. and Carpuat M. (2018). *Fluency over adequacy: A pilot study in measuring user trust in imperfect MT*. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, Boston, MA, pp. 13–25, Association for Machine Translation in the Americas.

Maruf S., **Saleh F. and Haffari G.** (2021). A survey on document-level neural machine translation: methods and evaluation. *ACM Computing Surveys* **54**(2), pp. 1–36.

Mathur N., **Wei J.**, **Freitag M.**, **Ma Q. and Bojar O.** (2020). *Results of the WMT20 metrics shared task*. In *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, pp. 688–725.

Melby A. K. and Foster C. (2010). Context in translation: Definition, access and teamwork. *Translation & Interpreting* **2**, pp. 1–15.

Rauf S. A. and Yvon F. (2022). Document level contexts for neural machine translation, Technical report, LIMSI-CNRS.

Scarton C., **Zampieri M.**, **Vela M.**, **van Genabith J. and Specia L.** (2015). *Searching for context: A study on document-level labels for translation quality estimation*. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, Antalya, Turkey, pp. 121–128.

Shirali-Shahreza S. and Penn G. (2023). Better replacement for TTS naturalness evaluation. In *12th Speech Synthesis Workshop (SSW) 2023*.

Tiedemann J. and Scherrer Y. (2017). *Neural machine translation with extended context*. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, Copenhagen, Denmark: Association for Computational Linguistics, pp. 82–92.

Toral A. (2020). *Reassessing claims of human parity and super-human performance in machine translation at WMT 2019*. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal, pp. 185–194, European Association for Machine Translation.

Toral A., **Castilho S.**, **Hu K. and Way A.** (2018). *Attaining the unattainable? reassessing claims of human parity in neural machine translation*. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium: Association for Computational Linguistics, pp. 113–123.

Virtanen P., **Gommers R.**, **Oliphant T. E.**, **Haberland M.**, **Reddy T.**, **Cournapeau D.**, **Burovski E.**, **Peterson P.**, **Weckesser W.**, **Bright J.**, **van der Walt S. J.**, **Brett M.**, **Wilson J.**, **Millman K. J.**, **Mayorov N.**, **Nelson A. R. J.**, **Jones E.**, **Kern R.**, **Larson E.**, **Carey C. J.**, **Polat İ.**, **Feng Y.**, **Moore E. W.**, **VanderPlas J.**, **Laxalde D.**, **Perktold J.**, **Cimrman R.**, **Henriksen I.**, **Quintero E. A.**, **Harris C. R.**, **Archibald A. M.**, **Ribeiro A. H.**, **Pedregosa F.**, **van Mulbregt P.**, **and SciPy 1.0 Contributors** (2020). SciPy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, **17**, 261–272.

Voita E., **Sennrich R. and Titov I.** (2019). *When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, pp. 1198–1212.

Voita E., **Serdyukov P.**, **Sennrich R. and Titov I.** (2018). *Context-aware neural machine translation learns anaphora resolution*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, pp. 1264–1274.

Wang L., **Tu Z.**, **Way A. and Liu Q.** (2017). *Exploiting cross-sentence context for neural machine translation*. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, pp. 2826–2831.

Cite this article: Knowles R and Lo C-k (2025). Calibration and context in human evaluation of machine translation. *Natural Language Processing* **31**, 1017–1041. https://doi.org/10.1017/nlp.2024.5