Error Span Annotation:
A Balanced Approach for Human Evaluation of Machine Translation

Tom Kocmi[★1]   **Vilém Zouhar**[★2]   **Eleftherios Avramidis**[3]   **Roman Grundkiewicz**[1]
Marzena Karpinska[4]   **Maja Popović**[5]   **Mrinmaya Sachan**[2]   **Mariya Shmatova**[6]

[1]Microsoft   [2]ETH Zurich   [3]DFKI
[4]UMass Amherst   [5]DCU & IU   [6]Dubformer

tomkocmi@microsoft.com   vzouhar@inf.ethz.ch

Abstract

High-quality Machine Translation (MT) evaluation relies heavily on human judgments. Comprehensive error classification methods, such as Multidimensional Quality Metrics (MQM), are expensive as they are time-consuming and can only be done by experts, whose availability may be limited especially for low-resource languages. On the other hand, just assigning overall scores, like Direct Assessment (DA), is simpler and faster and can be done by translators of any level, but is less reliable. In this paper, we introduce Error Span Annotation (ESA), a human evaluation protocol which combines the continuous rating of DA with the high-level error severity span marking of MQM. We validate ESA by comparing it to MQM and DA for 12 MT systems and one human reference translation (English to German) from WMT23. The results show that ESA offers faster and cheaper annotations than MQM at the same quality level, without the requirement of expensive MQM experts.

1   Introduction

While automatic evaluation metrics are important and invaluable tools for rapid development of Machine Translation (MT) systems, human assessment remains the gold standard of translation quality (Kocmi et al., 2023; Freitag et al., 2023). The translation quality is conceptually measured through adequacy (preservation of the original meaning) and fluency (grammaticality of the translated text; Koehn and Monz, 2006), and sometimes through comprehension (how readable or understandable the translation is; White et al., 1994).

Annotators are usually asked to assign a score on a particular quality aspect. Likert and 0–100 scale are often used for discrete and continuous scales.
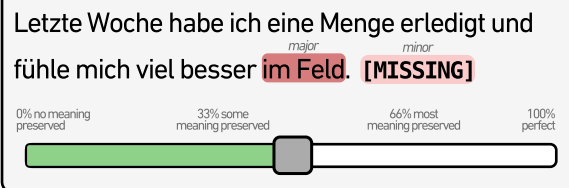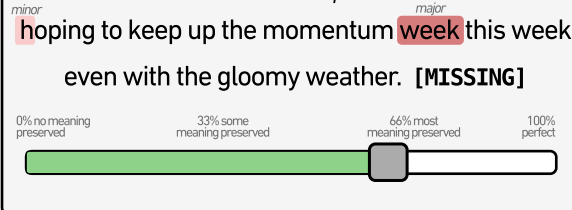

Figure 1: Stylized annotation user interface with Error Span Annotation (ESA). The annotator first marks errors with minor and major severity and then assigns a final score. This is more robust than asking for score directly.[1]

The most popular scoring method in machine translation field in recent years is Direct Assessment (DA; Graham et al., 2013), which is used to portray a human assessment of MT quality in the WMT shared tasks since 2016. Since 2022, the DA+SQM metric is used, namely direct assessment enriched with more objective Scalar Quality Metrics (SQM) guidelines (Kocmi et al., 2022).

Translation scores indicate the overall quality of a translation, but they can be subjective and do not provide details about the translation errors. The usual way to overcome this drawback is error classification: asking the evaluators to mark each translation error and assign an error tag from a set of predefined categories, such as *terminology* or *style*.

[★]Equal contributions. Others alphabetically.
[0]Code & collected data:
   github.com/wmt-conference/ErrorSpanAnnotation

---

[1]Our experiments are on English→German. In Figure 1, Spanish→English is only an illustration for English-speaking readers. The first example, based on our data, omits *Rainy days!* and incorrectly translates *about field items* as *im Feld* (=*in the field*). The second example, for illustrative purposes, does not capitalize *H* and mistakenly adds extra *week*.

In recent years, the dominant error classification protocol is the Multidimensional Quality Metrics (MQM; Lommel et al., 2014; Freitag et al., 2021a). MQM error classification is the standard human metric in the WMT Metrics shared task since 2021 (Freitag et al., 2021b). While error classification provides interesting insights into the distribution of different types of errors, it requires much more time and effort, both for annotators and task organizers, who need to prepare the error taxonomy, annotation guidelines and training examples.

We present a new evaluation protocol based on highlighting errors and followed by assigning scores, Error Span Annotation (ESA), and compare it to the MQM error classification and DA+SQM scores. We compare MQM, DA+SQM and ESA annotations in parallel on a subset of English→German machine translation outputs from WMT23. We find that the proposed ESA protocol is faster and cheaper than MQM whilst providing the same usefulness in ranking MT systems.

2 Related work

Assigning overall scores was the very first method of manual MT evaluation (ALPAC, 1966; White et al., 1994), where the evaluators assessed some or all of the translation quality criteria at once: adequacy, comprehensibility, and fluency. The first WMT (Workshop/Conference on Machine Translation) shared task in 2006 and the subsequent task in 2007 adopted this technique and used adequacy and fluency scores as official metrics (Koehn and Monz, 2006; Callison-Burch et al., 2007). Later, Vilar et al. (2007) proposed binary ranking of two or more MT outputs, which became the official metric at WMT 2008 (Callison-Burch et al., 2008). It required less effort and showed better inter-annotator agreement than adequacy and fluency scores. This remained the official WMT metric until 2017 when it was replaced by continuous Direct Assessment (DA). DA (Graham et al., 2013) does not use discrete scales, but a continuous one between 0 and 100. Bojar et al. (2016) scrutinized the quality criteria and recommended to focus on adequacy and use fluency to break ties only. DA replaced ranking methods in 2017 (Bojar et al., 2017) and since 2022 (Kocmi et al., 2022) it is used with SQM guidelines (Freitag et al., 2021a) in a slightly modified version with more descriptive scale labels, which increased the inter-annotator agreement.

None of the described methods provides information about the erroneous or problematic parts of the translation. An early work of Vilar et al. (2006) analyzes errors in translation outputs assigning them to error classes from a predefined error typology. Most popular error typology recently is Multidimensional Quality Metrics (MQM; Lommel et al., 2014; Klubička et al., 2018; Freitag et al., 2021a), which is used in WMT metrics task since 2021 (Freitag et al., 2021b).

Several error span marking methods have been proposed recently (Kreutzer et al., 2020; Popović, 2020) as a less demanding error annotation approach than error classification. While it does not provide the fine-grained details about different error classes, it still gives the information about the position and amount of errors, and also enables further fine-grained analysis on the annotated data, if necessary (e.g. classification of already marked errors, identifying linguistic phenomena causing the errors, or focusing on particular error type). While the previously reported findings on this method are promising, no systematic comparison for the purposes of evaluating machine translation systems has been carried out so far. Furthermore, the simplified error marking method does not solve the challenges in determining how to appropriately weight individual errors to obtain segment-level scores, a problem that becomes particularly pronounced when extending the evaluation to document level.

This work combines the advantages of error annotation (like MQM) and assigning direct scores (like DA). The annotators are first asked to identify and mark all errors, and afterwards to assign an overall score. When deciding about the score, they are *primed* by the preceding error annotation and see all the marked errors that can be taken into consideration for the final score.

3 Comparison with DA+SQM and MQM

Our proposed method lies between DA+SQM and MQM protocols, so we provide a detailed comparison between the two before describing ESA in details in the next section.

While both DA+SQM and MQM generally exhibit low inter-annotator agreement (Knowles and Lo, 2024; Freitag et al., 2021a), DA+SQM scores have high variance, which needs to be compensated with higher number of annotations per system (Wei et al., 2022). On the other hand, MQM requires human experts trained with the MQM protocol and

error classification. Trained experts can be twice as expensive as translators or bilingual speakers evaluating DA+SQM. The required expertise is a hard constraint which makes evaluation on some languages prohibitively expensive or not possible at all, especially low-resource languages. Furthermore, assigning a DA+SQM numerical score to a segment is anecdotally and intuitively much faster than MQM, where the evaluators need to mark each error span, classify it and assign severity. This altogether can make each MQM annotated segment up to approximately 10× more expensive.

DA+SQM is usually based on sentence-level scores, and the paragraph-level score is computed as the average of all sentence scores in the paragraph. Paragraph-level DA+SQM evaluation is possible, but evaluating an entire paragraph takes more time, substantially decreases the total number of collected scores, and is more demanding cognitively, which negatively impacts the inter-annotator agreement (Castilho, 2020). On the other hand, MQM as an error classification annotation is agnostic to the choice of annotation unit.

As an example, in App. Figure 6 we show the system ranking based on two approaches DA+SQM and MQM on Chinese-English (sentence-level evaluation) and English-German (paragraph-level evaluation) language pairs. Although both techniques reach same order of system clusters, DA+SQM produces much fewer clusters in paragraph-level setup, thus putting many systems within a single cluster. On the other hand, MQM is better able to distinguish different systems. To increase statistical power of DA+SQM, we would have to collect much more DA+SQM samples (Wei et al., 2022), which would further drive the cost up. In addition, DA is much more skewed towards fluency as opposed to adequacy (Martindale and Carpuat, 2018).

The cost difference was one of the main reason behind the high usage of DA+SQM at the WMT General MT shared task (Kocmi et al., 2023, 2022). For all these reasons, our hypothesis is that a new annotation protocol, ESA, which is between DA+SQM and MQM can provide better annotations than DA+SQM at a lower cost than MQM.

4 Error Span Annotation

Annotation process. In Error Span Annotation (ESA), the evaluators first mark all problematic parts (characters, words, phrases, sentences) in the translated text. For each marked span, they

Source/Translation+ESA	Score
SRC: ... I've entered the burrata dimension. **TGT**: ... ich habe die Burrata-Dimension eingegeben. **gloss**: *habe eingegeben(=I put in)* should be *bin eingetreten*	70%
SRC: Not like other tomb raider games **TGT**: Nicht wie andere Gräberüberfäller Spiele **gloss**: *Gräberüberfäller* overtranslates *Tomb Raider*	35%
SRC: (PERSON2) Yeah, so just know like- **TGT**: Ja, also weißt du einfach... [missing] **gloss**: *PERSON2 is missing*	86%
SRC: All collards, kale, chard is transplanted. **TGT**: Alle Kohlköpfe, Grünkohl, Schmalz sind verpflanzt. **gloss**: *Kohlköpfe(=cabbages)* and *Schmalz(=lard)* are incorrect translations	17%

Example 1: ESA-annotated examples with associated manual score. The error severity distinction is between minor and major. The first example has a single error (confusion *eingeben(=put in)* as *enter* meaning *go in*) which also affects the auxiliary verb habe/bin. Thus, the same error is marked twice.

also provide one of the two severity levels: **major** (e.g. changed meaning) or **minor** (e.g. incorrect grammar, style; see Example 1). Because all error spans are marked in the translation, not the source text, we include a special tag for marking omission errors. This was an intentional choice over annotating the source text to make the annotation protocol forward-compatible with other translation modalities, such as audio and video translations.

After the annotators mark all the error spans, they are asked to provide an overall score for the entire segment, on the scale from 0 to 100, reminiscent of DA+SQM. We implement the annotation interface in Appraise (Federmann, 2018) and show a screenshot in Figure 2. The full guidelines displayed to annotators are shown in Appendix A.

Segment-level scores. To rank systems, we need scalar values. There are two evident ways to extract them from ESA: (1) using the annotator's overall segment-level score directly, like DA+SQM, or (2) converting error span severity levels into a segment-level score, like MQM. By instructing the annotators to identify and mark all errors first, we *prime* them to be more accurate when assessing the overall quality of the segment—when making the decision about the score, they have already marked all errors in the segment and can see them, therefore they can take them into consideration.

MQM is primarily error diagnostics protocol which has been repurposed for translation segment scoring. The transition from MQM error spans

into a single segment score has been proposed to be done with the formula based on error severity counts (Freitag et al., 2021a):

$$\text{MQM-like} = -5 \cdot \#\text{MAJOR} - 1 \cdot \#\text{MINOR}$$

Notice that this does not scale with different text sizes.[2] As an example, if translation of a segment has two major errors, it receives the score of -10. However, if the source is repeated twice and the corresponding translation as well, the score would be further decreased to -20. This is especially problematic for paragraph-level evaluation which features segments of different length. Additionally, the segment-level MQM score might not correspond with the segment-level translation quality, such as when marking one error affecting several places in the segment, as in Example 1 (top). To avoid such issues, we use the annotators' direct scores as the main scoring approach for the ESA protocol unless specified differently. The system-level scores for all human evaluation methods in this work are calculated as the average of all segment-level scores for particular system. We further revisit the score computation in App. Appendix B.1. In some analysis, we use ESA error spans to calculate MQM-like score, we refer to such scores as ESA_spans.

Advantages. Assigning overall scores is guided by errors in the translation, and through error marking, the annotator can first focus on direct highlighting of these issues instead of determining the overall score directly. The advantage of ESA over error classification is that it is less demanding, while still informative—the annotations can be further refined in subsequent analyses. Furthermore, the evaluators are not limited to any pre-defined annotation protocol and can highlight a larger range of errors. The error marking approach can be seen as *descriptive* (encouraging annotator subjectivity and capturing their individual beliefs) and the error classification as a *prescriptive* (discouraging annotator subjectivity and asking annotators to align with one specific belief, in this case the pre-defined error protocol), as per Rottger et al. (2022).

5 Experimental setup

We conduct experiments comparing the ESA protocol with MQM and DA+SQM protocols. We design them in a way which makes it comparable with the previously collected annotations for

	ESA_1	ESA_2	MQM	MQM^WMT
# error spans	0.45	1.00	0.53	3.37
% minor	63%	68%	67%	67%
% major	37%	32%	33%	33%
Score (MQM-like)	81.8 (-1.1)	84.5 (-2.2)	(-1.2)	(-7.1)

Table 1: Average number of error spans per segment, ratio between minor and major errors, and scores across different annotation protocols.

MQM^WMT and DA+SQM^WMT. For this reason, we reproduce the human evaluation campaign for the WMT23 English to German systems (Kocmi et al., 2023; Freitag et al., 2023) with ESA and our reimplementation of MQM.

The original campaign featured 13 translations of 557 source segments. To facilitate running multiple campaigns for proper comparison, we had to scale down and subsampled 207 segments per system (74 documents), which yields 2,691 segments in total. To keep the ESA annotation comparable to other protocols, we subsample by selecting a subset of documents evaluated by Freitag et al. (2023) keeping the entire documents. This differs from Freitag et al. (2023), who removed some paragraphs from the ends of long documents. In our analysis we only consider segments overlapping with both previous annotation collections, thus obtaining 2,027 annotations evaluated across all protocols. We note that the subsampling makes all annotations protocols statistically less powerful, but keeps them fair in terms of statistical power per evaluated segment. Therefore, clustering and final system ranking in our analysis differs from Kocmi et al. (2023); Freitag et al. (2023). To keep the study comparable with Kocmi et al. (2023), we use the Wilcoxon rank-sum test with $p<0.05$ when producing system clusters. However, as all systems are evaluated on the same set of segments, we advice to use Wilcoxon signed-rank test when employing ESA as proposed by Kocmi et al. (2021).

To analyze inter annotator agreement, we run ESA protocol twice with different sets of annotators. We hired 28 annotators to evaluate our protocols and each protocol was evaluated by different sets of experts to avoid bias. Specifically, we had 8 bilingual annotators for the initial run of ESA_1, 10 translators for ESA_2 (different vendor), and 10 annotators to evaluate MQM protocol. For MQM, we hired professionals already experienced with MQM annotation protocol, while for ESA, we hired translators or bilingual speakers. All of them were native speakers of the target language, German.

Highlighting errors:

Highlight the text fragment where you have identified a translation error (drag or click start & end).

Click repeatedly on the highlighted fragment to increase its severity level or to remove the selection.

Minor Severity: Style/grammar/lexical choice could be better/more natural.

Major Severity: Seriously changed meaning, difficult to read, decreases usability.

If something is missing from the text, mark it as an error on the **[MISSING]** word.

The highlights do not have to have character-level precision. It's sufficient if you highlight the word or rough area where the error appears.

Each error should have a separate highlight.

Score: After highlighting all errors, please set the overall segment translation scores. The quality levels associated with numerical scores on the slider:

0: No meaning preserved: Nearly all information is lost in the translation.

33%: Some meaning preserved: Some of the meaning is preserved but significant parts are missing. The narrative is hard to follow due to errors. Grammar may be poor.

66%: Most meaning preserved and few grammar mistakes: The translation retains most of the meaning. It may have some grammar mistakes or minor inconsistencies.

100%: Perfect meaning and grammar: The meaning and grammar of the translation is completely consistent with the source.

Getting my eyes checked, because there was a decent deal for it here, and my left eye's been weird for some time now, so maybe they can tell me something what's happening.

Ich lasse meine Augen untersuchen, denn es gab hier ein gutes Angebot dafür, und mein linkes Auge ist seit einiger Zeit komisch, also vielleicht können sie mir sagen, was los ist. **[MISSING]**

| 0%: No meaning preserved | 33%: Some meaning preserved | 66%: Most meaning preserved | 100%: Perfect |
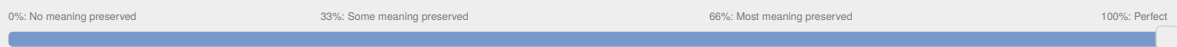
I'm splurging on a new set of frames, these red ones I reeeeally like.

Ich verschwende mein Geld für eine neue Brillenfassung, diese rote mag ich seeehr. **[MISSING]**

| 0%: No meaning preserved | 33%: Some meaning preserved | 66%: Most meaning preserved | 100%: Perfect |
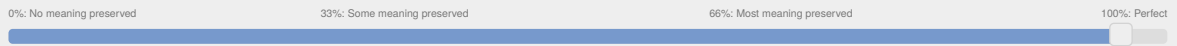
So, apparently the ghost image my left eye sees isn't too much of a concern. But I need basically varifocals too. So the lenses have a combo of those. Not the full on varifocals but something in between those and "normal" glasses (no idea of the terminology in English).

Also ist das Geistsrbild, das mein linkes Auge sieht, nicht so sehr Anlass zum Sorge. Aber grundsätzlich brauche ich auch Gleitsichtgläser. Die Linsen sind also eine Modifikation dieser. Nicht die starken Gleitsichtgläser, sondern etwa zwischen diesen und „normalen" Gläsern (keine Ahnung der Terminologie in Englisch). **[MISSING]**

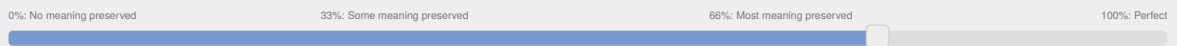| 0%: No meaning preserved | 33%: Some meaning preserved | 66%: Most meaning preserved | 100%: Perfect |

Figure 2: Screenshot of the beginning of one annotated document in the ESA interface (following segments are not shown). By showing and annotating whole documents at the segment-level, the annotators see all the relevant context. Segment *reset* button and *completed* labels removed for brevity. See the interactive tutorial shown to all annotators in Appendix Figure 10.

6 Analysis

We first analyze the collected data, system ranking, agreement with other protocols, quality assurance, and finally the annotation time. The findings reveal that the ESA quality is comparable, if not better than MQM, takes less time, and does not require highly trained annotators.

6.1 Score distribution

As per Table 1, on average the ESA_1 annotators mark 0.45 error spans per segments, which is close to MQM's 0.53 error spans per segment. The second run of ESA_2 has more than double of error spans per segment, which could be the result of different characteristics of annotators group, specifically experienced annotators in ESA_1 versus trans-

lators in ESA_2. On the other hand, MQM^{WMT} contains 7x more errors per segment than our re-run of MQM. We attribute this difference primarily to the differences in annotation crowds, which further motivates our own evaluation of both MQM and ESA so that the annotations differ only in the annotation protocol and past MQM training. The severity levels are distributed similarly across campaigns. Important insights are in the score range distribution presented in Figure 3: the MQM-like score computation creates more skewed distribution around 0, which is in addition unbounded and can go to -infinity the longer the evaluated segment is, complicating modeling and comparisons. In contrast, the manual scores from annotators are spread out and guaranteed to be in [0, 100].
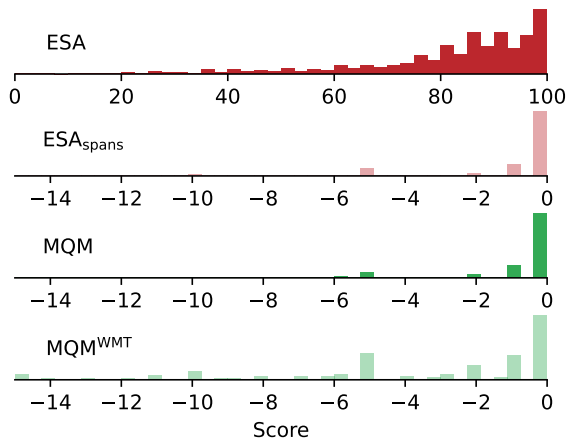
Figure 3: Distribution of scores for one annotation campaign. For ESA, we either use the manual score or ESA_{spans} computation based on error severities. For MQMs, the distribution is clipped ≥ -15 for higher resolution.

6.2 System ranking capabilities

We now investigate how well different annotation protocols (ESA and MQM) can rank MT systems. For purpose of this experiment, we consider MQM^{WMT} which comes from an independent very high-quality annotation crowd and implementation, as the gold standard. Note that this creates a positive bias towards our implementation of MQM as opposed to ESA. In Appendix B.2 we show an evaluation without this gold standard assumption.

When comparing two protocols, we ideally want them to rank all systems in the same order. This is not always possible as some systems are very similar and cannot be significantly distinguished with the evaluated sample size. Another problem is that different protocols may weight different phenomena (e.g. fluency or adequacy) differently. To compare different protocols in the task of ranking systems, we use pairwise accuracy (Kocmi et al., 2021), which is also used in WMT Metrics shared task when comparing different automatic metrics (Freitag et al., 2023). Pairwise accuracy measures how many system pairs does a protocol rank the same way as MQM^{WMT}. As we have only 78 system pairs, any wrong system pair will change pairwise accuracy by 1.28%. Therefore, we also calculate Spearman's correlations as we mainly want protocols to have monotonic ranking.

In Figure 4, each subplot compares system-level scores between one protocol on x-axis and MQM^{WMT} on y-axis. Our repeated MQM experiment and ESA protocol rank systems identically

(94.9%), while ESA has slighly higher Spearman's correlation with MQM^{WMT}. On the other hand, $DA+SQM^{WMT}$ significantly lacks behind both protocols. This suggest that our ESA protocol has comparable system ranking capabilities to MQM and is superior to $DA+SQM^{WMT}$.

The Figure 4 also shows, that relying on error spans only is not optimal, as ESA_{spans} has lower accuracy and Spearman's correlation than ESA. We can notice that this is even lower than our rerun of MQM. This can be attributed to the evaluation crowd, where we used professional MQM annotators for MQM protocol, while we used bilingual speakers and translators for the ESA protocol.

Further focusing on the clustering, MQM^{WMT} significantly differentiates the top system from others (highlighted in orange), while $DA+SQM^{WMT}$ strongly puts this system into the second cluster. This system is human reference, which we assume should be of highest quality. The reduced number of clusters in contrast to Figure 6 is due to the lower sample size. This may be result of $DA+SQM^{WMT}$ higher sensitivity to fluency and style errors, which contribute to 60% of all errors in human reference as marked by MQM^{WMT}. This conflict in clustering is one of the critiques of $DA+SQM^{WMT}$ if we assume that human reference should be the highest scoring translation. For example, in Kocmi et al. (2023), human reference was the best translation only in 2 out of 8 language pairs.

	MQM^{WMT}
ESA_1	0.227
$ESA_{1\ spans}$	0.170
ESA_2	0.250
$ESA_{2\ spans}$	0.236
MQM	0.189
$DA+SQM^{WMT}$	0.209

Table 2: Kendall τ segment-level correlations between evaluation protocols.

6.3 Agreement with other protocols

We now compare how different protocols correspond on the segment-level. We use MQM^{WMT} as the gold standard to compare against because it was done independently outside of our setup and with high quality assurance. We analyze two aspects: (1) segment scores, and (2) spans, where we consider spans overlapping even with a single
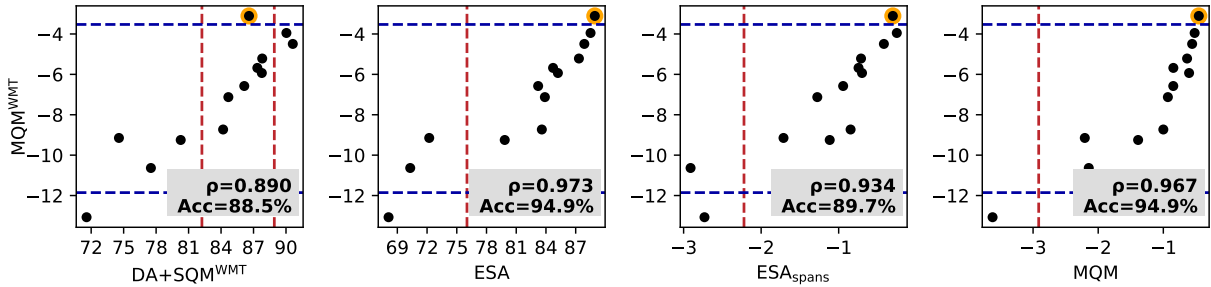
Figure 4: Each point represents a system, with the original MQM$^{\text{WMT}}$ scores on the *y*-axis plotted against our rerun of DA+SQM$^{\text{WMT}}$ (first plot), ESA (second plot), ESA$_{\text{spans}}$ (third plot), and MQM (forth plot). Stripped lines indicate cluster separations determined by each method with alpha threshold 0.05. We compute Spearman correlation ρ and pairwise accuracy Acc.

character as a match. For this evaluation, we use Kendall τ variant C, which is more suitable for data with different underlying scales and many ties.

In Table 2, we see that although all protocols correlate similarly with MQM$^{\text{WMT}}$, our protocols obtain the highest τ for both runs. The segment-level correlation also confirms that relying only on the error spans is not optimal and ESA$_{\text{spans}}$ obtains lower Kendall score.

In Table 3 we focus on the annotated spans. We consider any spans that overlap as matching, irrespective of severity. Because different protocols have different average number of error spans, presenting just the size of the intersection would be misleading. Instead, we show normalized set similarity that is not symmetric. It answers the questions: *What proportion of samples in B were covered by A?* Both ESA and MQM cover MQM$^{\text{WMT}}$ similarly, with 29% and 32% respectively. At the same time, MQM$^{\text{WMT}}$ covers ESA and MQM with 93% and 85%.

↓A B→	ESA	MQM	MQM$^{\text{WMT}}$
ESA		77%	29%
MQM	85%		32%
MQM$^{\text{WMT}}$	93%	89%	
$|A \cap B|/|B|$			(100% for $B = \emptyset$)

Table 3: Similarity between spans of different annotation protocols (one campaign) computed as percentage how much of B does A contain. For example, 93% of ESA spans were also in MQM. Any span that overlaps with another one is considered a hit, even if only with as single character.

6.4 Quality of annotations

Intra annotator agreement. We want the human evaluation protocol to consistently assign similar scores for the same translations over time. A good

indicator of the annotation quality is how noisy and subjective it is, which is reflected by how much annotators agree on the same segments (inter annotator agreement) as well as how a single annotator agrees with themselves (intra annotator agreement). To measure intra-AA, we ask the same annotators to again annotate the same documents two months later. We prepare an identical campaign with the same distribution of systems in the same order as originally, asking the same annotators to redo it again for both ESA and MQM.

It is not obvious how to measure the agreements for protocols that have different features. One issue is the frequency of ties, where MQM has more ties than rating from ESA or DA+SQM$^{\text{WMT}}$. For example, MQM$^{\text{WMT}}$ contains 30.8% no-errors, while DA+SQM$^{\text{WMT}}$ contains only 5.2% of score 100. Secondly, each protocol uses different range and distribution of scores, which makes calculation of agreement complicated. See App. Figure 7 to understand different distribution of scores.

Previous works comparing different protocols discretize the scale into bins (Graham et al., 2013; Freitag et al., 2021a), however, this approach is sensitive to subjective selection of bin sizes and benefits already discrete protocols. Instead, we propose to use Kendall's Tau-c correlation to measure inter-annotator agreement. Secondly, we also want to take into consideration that small changes in scores are less damaging than large shifts, therefore, we want inter annotator's scores to correlate linearly, ideally having identical score each time. To measure this, we use Pearson's correlation. Lastly, we measure recall of how often annotator mark *any* error in the same segment in contrast to leaving the segment without marked errors.

Table 4 shows that ESA has all scores higher than MQM. Higher Kendall and Pearson suggest that the task is easier for annotators to agree on the

score. We expect the recall to be comparable for both techniques as the task is similar, we hypothesize that the drop in MQM could be explained by annotators saving time and skipping minor errors as the annotation is more complicated for MQM than ESA, which can be confirmed when looking at minor error's recall only.

Lastly, evaluating inter annotator agreement is heavily affected by the strategy of annotators, where different annotation strategy does not mean different performance of the task as Riley et al. (2024) showed. Secondly, the MQM^{WMT} was collected with a different tooling and the documents have been presented to annotators in different order, which could also impact the inter annotator agreement.

	Intra AA		Inter AA	
	ESA	MQM	ESA	MQM
Kendall's Tau-c	0.149	0.109	0.254	0.116
Pearson	0.403	0.189	0.482	0.281
Error recall	69.6%	61.9%	66.6%	40.1%
Minor e. recall	70.7%	66.2%	67.7%	44.4%
Major e. recall	82.6%	82.1%	84.8%	62.9%

Table 4: Intra- and inter-annotator agreement on segment-level.

Inter annotator agreement. To measure how different annotators agree between themselves on the same protocol, we compare ESA_1 to ESA_2, where each protocol was evaluated by different group of annotators (bilingual annotators vs. translators). To calculate MQM's inter-annotator agreement, we compare our MQM run with MQM^{WMT}. However, the comparison is not as 1:1 as for ESA protocol. Our MQM was collected with different interface and system outputs have been shown to annotators in a different order. Results in Table 4 suggest, that ESA has also higher inter-annotator agreement than MQM. Unfortunately, we could not rerun DA+SQM protocol to calculate agreements, which needs to be reevaluated in future work.

Agreement on the error span. We now investigate how much MQM annotators agree on the error spans, error categories and severity levels. We evaluate from two angles: *intra*, where we check if the same annotator marks the same error spans or have overlapping parts, and *inter*, where we compare our MQM error spans to MQM^{WMT}. Table 5 shows that only in 30% of cases, the same annotator marked at least part of the same segment as an

error regardless error severity and category. If we look at cases preserving severity and category, this number drops to 8.4%.

When comparing the inter annotator agreement, only 50% of errors are overlapping. This number is higher as the total number of errors in MQM^{WMT} is $7\times$ higher than in MQM, therefore it is more likely an error will have overlap.

	Intra AA	Inter AA
Any errors	29.3%	50.2%
Same severity	16.9%	23.7%
Same category	18.9%	24.1%
Same sev. + categ.	11.6%	10.0%
Same sev. + subcateg.	8.4%	-

Table 5: ESA intra- and inter-annotator agreement (frequency) on marking overlapping errors with same severities, categories or subcategories.

Quality control. To measure the quality of annotations, we added "attention checks" in the form of segments for which we can reliably check whether the annotator annotated them correctly or not. In random documents, we perturbed the translation by replacing part of it with random sequence of words of the same length introducing a major translation error. Within 100 segments annotated by the annotator, they see both original and perturbed versions of that document. We can control the annotation quality by checking if the perturbed documents received more error spans. We show a worked-out Example 2.

SRC*: Sie haben gestern das Treffen wieder verschoben.*
TGT*: He postponed the meeting again yesterday.*
TGT^P*: He postponed the meeting <u>squirrels tense</u>.*

Example 2: An example of a perturbed translation **TGT^P** based on the original system translation **TGT**. The **TGT** has one error (*He* should be *You* or *They*) and **TGT^P** introduces one more errors (*squirrels tense*).

The MQM and ESA setups used the same perturbations and we show their results in Table 6. The scores comparing original and perturbed segments for ESA and MQM are vastly different,[3] showing that annotators paid attention to the quality control items. For ESA the original segment had a higher score than the perturbed one in 86% of cases, while for MQM in 78% of cases. When investigating

---

[3]The protocols use different scales, for example, one major error under MQM is -5 points, while 25 points in ESA represents quarter of the full scale.

whether an annotator marked the error span or not, MQM has higher recall than ESA, however, this is less crucial for ESA as annotator can adjust the ranking without marking the error.

		Original	Perturbed	OK
ESA	Score	79.5	52.6	86%
	Span count	0.85	1.86	54%
	Perturbation marked			56%
MQM	Score	-1.87	-6.49	78%
	Span count	0.66	1.68	70%
	Perturbation marked			76%

Table 6: Annotations assigned to perturbed attention check items (either scores or number of spans). **OK** is percentage in how many cases the non-perturbed item received a higher score or had fewer error spans, and how often the perturbed span was marked by the annotator.

6.5 Annotation time

One of the reasons behind development of the new protocol is reducing the time requirements of the evaluation. In this section, we analyze times on our experiments only, therefore our rerun of MQM in contrast to ESA. We do not include the times for DA+SQM and MQM^{WMT} because they were done independently outside of our study and the time data are not available. Assessing the speed of annotations is challenging as annotators took breaks during the annotation (from short breaks taking several minutes up to several hours). This makes the evaluation of time problematic and we therefore investigate the time estimate in several ways. Counting all annotators together, the median time for annotation per single paragraph for MQM is 38 seconds and for ESA is 29 seconds, a reduction of 23%. However, as median time for each annotator fluctuates, we look at the average median time across annotators. **For MQM, the median is 49 seconds and for ESA it is 34 seconds, a reduction of 32%**. This can be contributed mainly to the less demanding error span annotation approach used in ESA.

We note that further speedups could be made by instructing the ESA annotators to not spend extra time e.g. marking multiple error-span annotations of a single grammatical phenomenon.

Speedup during annotations. Naturally, the reported total annotation time does not distinguish between the duration of the first and last annotations. In practice, annotators *learn* to perform the

annotation task more effectively. In Figure 5 we show time per segment depending on how many segments the annotator already processed. For both MQM and ESA, there is a small learning effect. For MQM, with each segment, the annotator becomes 0.20s faster, while for ESA this is 0.17s.
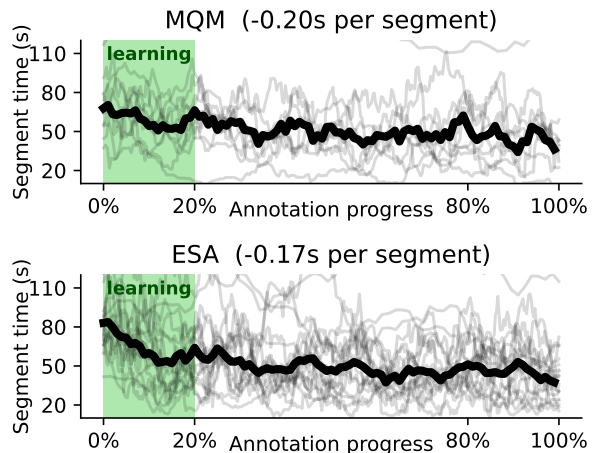

Figure 5: Time per segment with respect to progression in the annotation. The faint gray lines represent individual annotators, while the bold black line shows the average time. The lines are smoothed with a window of size 15 segments. We also compute the average speed at the beginning and at the end, which yields the *learned speedup*. This is how much the annotator speeds up after working on one segment.

7 Conclusion

Existing annotation protocols for machine translation evaluation are either expensive because they require expert labor (MQM), or they are noisy and less reliable (DA+SQM). To this end, we propose, describe, and analyze **Error Span Annotation (ESA)**, which builds on top of previous protocols to enable economic evaluation at scale. It works by asking the annotators to mark error spans, but with only the error severities and not types. In contrast to MQM, we also solicit final translation score. This is more reliable than DA+SQM, as the annotators are primed and informed about the translation errors to assess quality of longer documents.

We showed that our protocol has the higher inter and intra annotator agreement than MQM while being 32% faster. In addition, the protocol does not require annotators trained in MQM categorizations.

Lastly, we showed that relying only on error spans and not using the ranking score as we did in ESA$_{spans}$ produces suboptimal scoring, therefore the combination of error spans and ranking seems to produce the best results.

Limitations

A possible limitation in contrast to MQM is that now the system evaluation does not provide breakdown of error types, which could help practitioners in improving their systems. ESA does not provide this because the goal is *evaluation* and not *diagnosis*. Because of this and the costs of scaling expert labor, we are convinced that this is not a true shortcoming of ESA. Furthermore, the annotated errors can be further classified and analysed, if necessary.

Our experiments, for monetary reasons, were done only on one language pair, English→German. Nevertheless, it is unlikely that the results would be vastly different for other languages. The most difficult setup could be with Chinese, Japanese, and Korean texts that do not use spaces. However, we made a deliberate decision to allow highlighting of individual characters, as opposed to words, so that the user experience is unified across all languages. This was done in spite of speed improvements (selecting on the word level is easier than selecting individual character boundaries) in order to make the tool scalable to a large range of languages.

Ethics Statement

The annotators were paid a standard commercial translator wage in the respective country. The experts in the MQM annotation has been paid double the hourly wage. No personal data was collected and the showed data was screened for potentially disturbing content.

We follow up with a questionnaire asking annotators on their feedback. Almost all annotators specified that the annotation experience was positive and instructions were clear. The main concern they mentioned was that some documents have been too long to evaluate.

References

ALPAC. 1966. Language and machines. Computers in translation and linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, 169–214. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, An-tonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 131–198. Association for Computational Linguistics.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*. Aslib.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, 136–158. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, 70–106. Association for Computational Linguistics.

Sheila Castilho. 2020. 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*, 1150–1159. Association for Computational Linguistics.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 86–88. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, 578–628. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human

evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, 733–774. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 33–41. Association for Computational Linguistics.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2018. Quantitative Fine-grained Human Evaluation of Machine Translation Systems: A Case Study on English to Croatian. *Machine Translation*, 32(3):195–215.

Rebecca Knowles and Chi-kiu Lo. 2024. Calibration and context in human evaluation of machine translation. *Natural Language Processing*, 1–25.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, 1–42. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 1–45. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, 478–494. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, 102–121. Association for Computational Linguistics.

Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. 2020. Correct me if you can: Learning from error corrections and markings. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 135–144. European Association for Machine Translation.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, 165–172. European Association for Machine Translation.

Marianna Martindale and Marine Carpuat. 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)*, 13–25. Association for Machine Translation in the Americas.

Maja Popović. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5059–5069. International Committee on Computational Linguistics.

Parker Riley, Daniel Deutsch, George Foster, Viresh Ratnakar, Ali Dabirmoghaddam, and Markus Freitag. 2024. Finding replicable human evaluations via stable ranking probability. *arXiv preprint arXiv:2404.01474*.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 175–190. Association for Computational Linguistics.

David Vilar, Gregor Leusch, Hermann Ney, and Rafael E. Banchs. 2007. Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation*, 96–103. Association for Computational Linguistics.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA).

Johnny Wei, Tom Kocmi, and Christian Federmann. 2022. Searching for a higher power in the human evaluation of MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 129–139. Association for Computational Linguistics.

John S. White, Theresa A. O'Connell, and Francis E. O'Mara. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2024. AI-assisted human evaluation of machine translation.

A User Guidelines

The following are annotation guidelines for our local ESA and MQM campaigns.

A.1 ESA (Error Span Annotations)

Higlighting errors: Highlight the text fragment where you have identified a translation error (drag or click start & end). Click repeatedly on the highlighted fragment to increase its severity level or to remove the selection.

- **Minor Severity:** Style/grammar/lexical choice could be better/more natural.
- **Major Severity:** Seriously changed meaning, difficult to read, decreases usability.

If something is missing from the text, mark it as an error on the **[MISSING]** word. The highlights do not have to have character-level precision. It's sufficient if you highlight the word or rough area where the error appears. Each error should have a separate highlight.

Score: After highlighting all errors, please set the overall segment translation scores. The quality levels associated with numerical scores on the slider:

- **0%**: No meaning preserved: Nearly all information is lost in the translation.
- **33%**: Some meaning preserved: Some of the meaning is preserved but significant parts are missing. The narrative is hard to follow due to errors. Grammar may be poor.
- **66%**: Most meaning preserved and few grammar mistakes: The translation retains most of the meaning. It may have some grammar mistakes or minor inconsistencies.
- **100%**: Perfect meaning and grammar: The meaning and grammar of the translation is completely consistent with the source.

A.2 MQM (Multidimensional Quality Metrics)

Higlighting errors: Highlight the text fragment where you have identified a translation error (drag or click start & end). Click repeatedly on the highlighted fragment to increase its severity level or to remove the selection.

- **Minor Severity**: Style/grammar/lexical choice could be better/more natural.
- **Major Severity**: Seriously changed meaning, difficult to read, decreases usability.

If something is missing from the text, mark it as an error on the **[MISSING]** word. The highlights do not have to have character-level precision. It's sufficient if you highlight the word or rough area where the error appears. Each error should have a separate highlight.

Error types: After highlighting an error fragment, you will be asked to select the specific error type (main category and subcategory). If you are unsure about which errors fall under which categories, please consult the typology definitions.
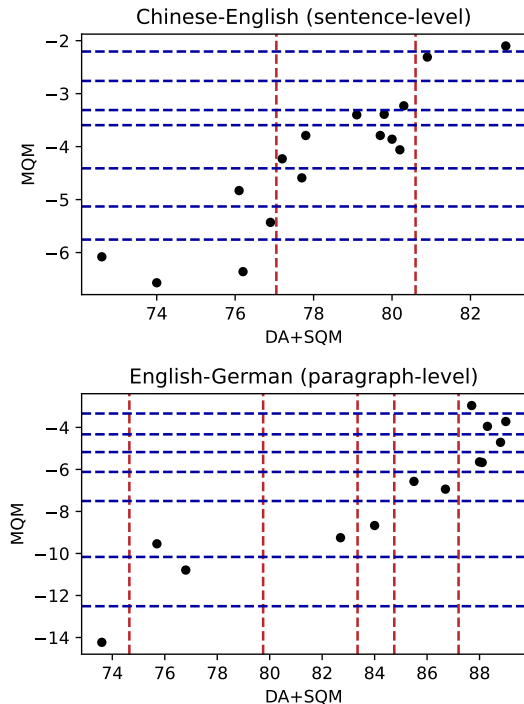

Figure 6: System scores illustrating differences between DA+SQM and MQM. Each point is a single system and dashed lines mark clusters. DA+SQM produces fewer clusters and groups many systems into one single cluster, while MQM better distinguishes different systems. Scores and clusters are from Kocmi et al. (2023).
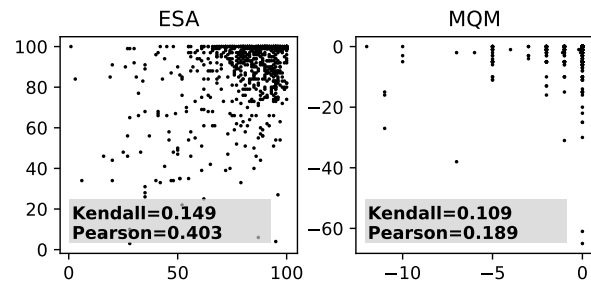

Figure 7: Intra annotator agreement; changes in scoring by the same annotator when evaluated again. Each point represents an annotated segment with x-axis being annotator's score assigned in March and y-axis their score assigned in May.

Feature	Corr. with ESA score (ρ)
Source token count	-0.16
Target token count	-0.06
Minor error count	-0.20
Major error count	-0.52
Missing error count	-0.45
Minor error count (normalized)	-0.13
Major error count (normalized)	-0.37
Missing error count (normalized)	-0.31

Table 7: Segment-level Pearson correlation of individual features with the ESA score.

B Additional results

B.1 From error spans to final score

To find out what influences the score, we show correlation between individual segment-level features in Table 7. On average, longer segments have lower translation quality. Importantly, the error counts normalized by segment length correlate less than the non-normalized counterparts. However, we note that the normalized scores are more continuous that the non-normalized MQM computation, as per Figure 3.

The MQM formula was crafted with respect to preserving system ranking and not segment-level matching from the same annotator (Freitag et al., 2021a). However, the construction of ESA allows us to revisit this problem as each annotator gives both the error spans and the final score. We scan for multiple minor/major ratios of error weights and show the results in Figure 8. We find that the optimal formula that optimizes the correlation between the direct score and the score from the spans has the following form: $\text{SEG.SCORE} = -1 \cdot \text{\#MINOR} - 4.8 \cdot \text{\#MAJOR}$, which is very close to the originally proposed 1:5 ratio. From Figure 8, the weight of the major error class seems to have a much bigger effect on the final translation score, suggesting that minor errors play a lesser role.
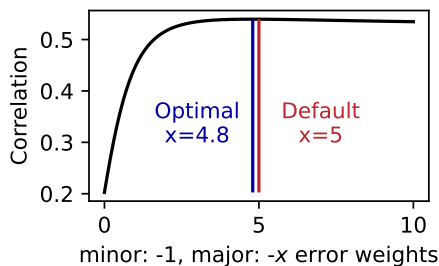

Figure 8: Correlation between direct ESA score and scores computed from error spans with minor errors having weight -1 and major errors $-x$.

B.2 Protocol evaluation without a gold standard

Evaluating the quality of a protocol without a target to compare to is difficult. In previous sections we assumed that MQM^{WMT} is the gold standard, which might bias the evaluation in favor of MQM. Even though the results showed higher correlations of ESA with MQM^{WMT}. For completeness, we consider the methodology of Zouhar et al. (2024)

which does not require target gold standard to compare the quality of annotation protocols.

The assumption is that annotation protocols have various levels of noise, but are unbiased in what they measure. Because MQM and DA+SQM might measure different things, we want to compare each to the perfect ranking of the particular thing they aim to measure. The linking hypothesis is that even noisy and low-quality annotation protocols would lead to the final system ranking with large enough data. Vice-versa, only robust annotation protocols would arrive at the final ranking with only a few data points per each system. This is formalized by the subset consistency accuracy. It is the system ranking accuracy on a subset of annotations with respect to the ranking induced by the full data from one annotation protocol.

We show the results in Figure 9. Out of the comparable lines, ESA (with direct scoring) achieves the highest subset consistency accuracy. In practice, this translates to needing fewer annotated examples to achieve the final system ranking. This directly corresponds to lower annotation costs.
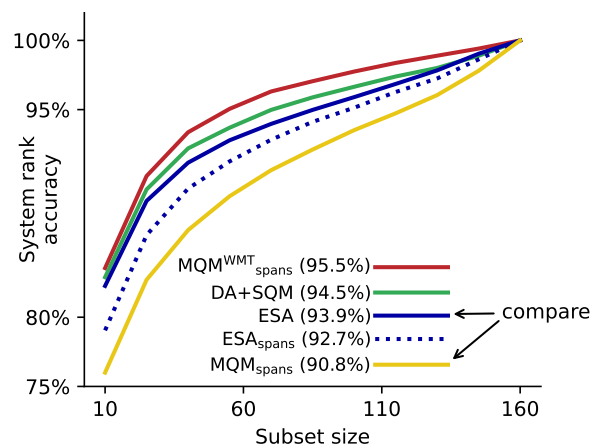

Figure 9: Subset consistency accuracy (Zouhar et al., 2024) of annotation protocols. E.g. with just 60 annotated segments, MQM^{WMT} achieves 95% system ranking accuracy with its final ranking based on 160 annotated segments. Values in the legend are averages, corresponding to normalized area under the curve. The only comparable lines are ESA, $\text{ESA}_{\text{spans}}$, and $\text{MQM}_{\text{spans}}$ because they were run in the same setting with similar crowds.

TUTORIAL: This translation seems to be all correct. Please use the slider to set the quality to 100%.

Der Hund ist rausgerannt.

The dog ran outside. **[MISSING]**

0%: No meaning preserved   33%: Some meaning preserved   66%: Most meaning preserved   100%: Perfect

Reset   ✓ Completed

TUTORIAL: The word "walked" is incorrect. Mark it minor error (light pink) using the instructions above and move the slider to 80%.

Der Hund ist rausgerannt.

The dog walked outside. **[MISSING]**

0%: No meaning preserved   33%: Some meaning preserved   66%: Most meaning preserved   100%: Perfect

Reset   ✓ Completed

TUTORIAL: The words "stayed inside" are very wrong. Mark them using the instructions and raise its severity to major. Then move the slider to 20%.

Der Hund ist rausgerannt.

The dog stayed inside. **[MISSING]**

0%: No meaning preserved   33%: Some meaning preserved   66%: Most meaning preserved   100%: Perfect

Reset   ✓ Completed

TUTORIAL: While the translation is technically correct, it does not sound very natural. Don't mark any erroneous words but at the end, use the slider to evaluate the overall translation quality as around 70%.

Although the cats stayed outside overnight, they were not cold.

Obwohl die Katzen die Nacht über im Freien verharrten, erfuhren sie keine Kälte. **[MISSING]**

0%: No meaning preserved   33%: Some meaning preserved   66%: Most meaning preserved   100%: Perfect

Reset   ✓ Completed

TUTORIAL: The word "Hund" (dog) is not present in the translation. Mark the [MISSING] text with major severity. Move the slider to 5%.

Der Hund ist rausgerannt.

The walked outside. **[MISSING]**

0%: No meaning preserved   33%: Some meaning preserved   66%: Most meaning preserved   100%: Perfect

Reset   ✓ Completed

TUTORIAL: The word "ran" is mistakenly marked as incorrect. Fix it by removing it according to the instructions. Move the slider to 100%.

Der Hund ist rausgerannt.

The dog ran outside. **[MISSING]**

0%: No meaning preserved   33%: Some meaning preserved   66%: Most meaning preserved   100%: Perfect

Reset   ✓ Completed

Figure 10: Tutorial to ESA annotations shown at the beginning of the campaign. All tutorial segments need to be annotated correctly before continuing.