Analysis of the Annotations from a Crowd MT Evaluation Initiative: Case Study for the Spanish-Basque Pair

Nora Aranberri
HiTZ Basque Center for Language Technologies - Ixa NLP Group
University of the Basque Country (UPV/EHU)
nora.aranberri@ehu.eus

Abstract

With the advent and success of trainable automatic evaluation metrics, creating annotated machine translation evaluation data sets is increasingly relevant. However, for low-resource languages, gathering such data can be challenging and further insights into evaluation design for opportunistic scenarios are necessary. In this work we explore an evaluation initiative that targets the Spanish—-Basque language pair to study the impact of design decisions and the reliability of volunteer contributions. To do that, we compare the work carried out by volunteers and a translation professional in terms of evaluation results and evaluator agreement and examine the control measures used to ensure reliability. Results show similar behaviour regarding general quality assessment but underscore the need for more informative working environments to make evaluation processes more reliable as well as the need for carefully crafted control cases.

1 Introduction

Particularly since trainable neural automatic metrics took centre stage in the WMT metrics shared task in 2022 (Freitag et al., 2022) machine translation (MT) evaluation data sets annotated for quality are becoming essential to develop accurate models. If availing of parallel data with professional references was not difficult enough, we are currently faced with the need to collect data that is not otherwise produced, in other words, while parallel data could be gathered from previously published translations, a sentence-(or text-)level numeric quality assessment most usually needs to be generated for the specific task of metric training.

This situation poses a particular challenge for low-resource languages which widens the gap between high- and low-resource scenarios. Firstly, because pre-trained models such as COMET (Rei et al., 2020) might not include language-specific data for small languages and therefore quality predictions can be unreliable, and secondly, because collecting relevant annotated data requires a heavy investment. In this context, resorting to opportunistic data collections with crowd volunteers is increasingly tempting.

This new scenario is yet an added reason to increase research efforts on evaluation design. More rigorous considerations of evaluation methodologies and design decisions emerged with claims of human and super-human parity of MT performance (Hassan et al., 2018; Barrault et al., 2019). Researchers claimed that evaluations were not rigorous and pointed out issues such as raters' lack of translation expertise, the quality of reference translations, target language interference in source sentences and non-contextualised evaluations as aspects that skewed results in favour of MT contenders (Läubli et al., 2018; Toral et al., 2018).

Reports of large evaluation initiatives and third-party reviews have shown that little by little evaluation approaches take into account some considerations (Toral, 2020; Popel et al., 2020) and reference campaign such as the annual WMT share task have taken steps to follow best practices for reliable evaluations (Kocmi et al., 2023). Adding to this, research on design-related topics are emerging, such as error methods to optimise test set configuration to reduce evaluation effort (Saldías Fuentes et al., 2022), classification schemes adapted to identifying critical errors in neural MT (Sudoh et al., 2021), document-level and context-aware agreement and effort (Castilho et al., 2020; Castilho, 2021), detection of post-edited reference translations (Kloudová et al., 2021) and differences between expert and non-expert evaluators (Graham et al., 2013; Freitag et al., 2021). Several crowd evaluation initiatives have also been reported over the years (Bentivogli et al., 2011; Graham et al., 2017) even for low-resource languages (Aranberri et al., 2017; Toral et al., 2017) that cover a number of design decisions. And yet, best practice and efficiency recommendation guidelines are limited and it is not uncommon that evaluation initiatives specially for low-resource scenarios lack the rigour that would benefit the outcomes the most. In this context, the current analysis is only a small step towards studying the characteristics of crowd-based evaluations within minority language communities.

The remaining of this paper is divided as follows Section 2 provides a brief description of the evaluation set-up from where the data set under study originated together with the obtained results; Section 3 outlines the specific details of the evaluation set-up used to obtain a professional evaluation of the said set as well as the qualitative feedback collected on the task and reports a comparison of evaluation results and the agreement analysis between the crowd volunteers and the professional evaluator; Section 4 examines the reliability of the control measures included in the set to identify outlier evaluators; finally, Section 5 draws a number of conclusions from the study.

2 Description of the Original Evaluation Initiative

The data set studied in this work is the product of an evaluation initiative to obtain human assessments of MT for two low-resource languages, namely, Basque and Maltese (Falcão et al., 2024). The authors aimed to collect sentence-level direct assessments to test the potential improvement of the trainable COMET metric with language-specific data. The resulting data set for the Spanish–Basque pair was kindly made available by the researchers for further analysis.[1]

[1]Access to the data set will be open upon publication of their work.

In this section, we briefly describe the evaluation setup used by the original research (for further details, see Falcão et al. (2024)) and report the overall results for later comparison.

2.1 Evaluation Set-up

Dataset: The evaluation set prepared for the campaign consisted of 400 Spanish source sentences and Basque translations. They were extracted from various existing sets and sources such as FLORES-200[2], TED2020 (Reimers and Gurevych, 2020), OpenSubtitles (Lison and Tiedemann, 2016), the Elhuyar Corpus[3] and the HAC parallel corpus[4], which cover text from web articles to subtitles and literature. Note that the Spanish source sentences in these sets can include both original and translated text.

Translation sources: The Basque translations paired with the Spanish sentences were obtained from multiple sources. Three MT systems were used to translate the set automatically. Additionally, damaged translations –MT system outputs with an embedded Spanish sequence of words- and reference translations –obtained from the parallel data sets– were also included in the final set as a means to identify unreliable evaluators.

Task: Distributed through the Appraise[5] platform (Federmann, 2012), the task involved evaluators assessing the translation quality in a continuous scale of 0 to 100. Directed towards a non-specialist participant profile, the description of the task highlighted a series of attributes, including meaning, information, clarity, correctness, grammaticality, and naturalness.[6] It could be argued

[2]https://github.com/facebookresearch/flores/blob/main/flores200/README.md
[3]https://elhuyar.eus/en/services/language-services-and-basque-plan/translation-and-language-resources/corpus
[4]https://www.ehu.eus/ehg/hac
[5]https://github.com/cfedermann/Appraise/ https://github.com/AppraiseDev/Appraise
[6]The original English text provided in the platform in the relevant languages was as follows: "For each item, you will be shown an original sentence in Spanish and a translation candidate in Basque. You will then be asked to rate the quality of the translation on a scale of 0 to 100, based on how well you believe the translation expresses the full meaning of the original sentence. A rating of 100 means that the candidate is a perfect translation: it expresses the same thing as the original sentence, in a clear and correct manner. A candidate should be rated lower if it contains grammatical or orthographic errors, if it's missing information, if it sounds unnatural or weird, and so on." (personal communication, J. Falcão, July 2023)

that the explanation aimed for a general definition of quality rather than a specific aspect. To perform the assessment, evaluators were provided with a source sentence and its corresponding translation. The sentences were provided without context. Participants were free to annotate as many pairs as they wished. No further guidelines were provided as to how to perform the task.

Evaluators: Crowd volunteers were sought by promoting the initiative through university and translator distribution lists, and social media. Therefore, the linguistic profiles of potential evaluators ranges from professional translators to general users with no dedicated training in languages. The evaluators were asked to report their Spanish and Basque language competence to exclude those without an advanced level of both languages. None such cases were reported by the researchers.

2.2 Evaluation Results

A quick analysis of the metadata reveals that 44 crowd volunteers contributed with a total of 1,186 evaluations (an average of 26.95 evaluations per person, with a median of 11. As shown in Table 1, their work is divided as follows: a total of 742 sentence pairs were evaluated,[7] 389 (%52.42) of which were assessed once, 285 (%38.41) twice and 76 (%10.24) received between three and five annotations. This allowed to collect annotations for about 200 sentences for each MT system (MT1, MT2, MT3), a total of 78 damaged translations, about 25 for each brand of damaged cases (D-MT1, D-MT2, D-MT3), and 53 sentence pairs containing reference translations (Ref).

According to the annotations, MT1 and MT2 score very similarly with results of 77.81 and 78.45 points, respectively (see Table 2).[8] MT3 lags behind, over 16 points lower. As anticipated, damaged translations score considerably lower, yet following the ranking for the MT systems. Unexpectedly, reference translations score lower than the system averages. As a general trend, the average scores tend to be higher for sentence pairs with a single annotation than for those with multiple annotations. These comparisons should be taken with caution as the sentences annotated for each subgroup are not exactly the same.

[7]Note that this does not cover the over 1,200 pairs in the evaluation set.

[8]Scores for sentence pairs with more than one evaluation were calculated separately; no average was applied.

		Evaluations			
Trans. source	**Sentences**	**1**	**2**	**+2**	**Total**
MT1	213	112	78	23	341
MT2	207	97	89	21	342
MT3	191	112	64	15	286
D-MT1	28	14	9	5	48
D-MT2	26	13	10	3	44
D-MT3	24	15	5	4	39
Ref	53	26	22	5	86
Total	742	389	285	76	1186

Table 1: Number of evaluated sentences and collected evaluations, where Trans. source refers to the source from where the translations were obtained, Sentences refers to the number of unique sentence pairs assessed, and evaluations 1, 2 and +2 refer to the number of sentences that obtained the stated number of evaluations.

3 Professional Evaluation

In order to explore the similarity between crowd and professional evaluators and their reliability, the author performed the same evaluation task for the complete set. She is a specialist in translation, native speaker of Basque (accredited C2-level) and Spanish, and has experience in MT evaluation. She will be referred to as the professional evaluator. While the feedback from a single professional is not necessarily indicative of the true annotations the sentence pairs should receive, it can be argued that it provides an educated guess that is consistent across the set to the extent this is possible in human judgement.

3.1 Evaluation Set-up

As in the original evaluation, the evaluator was presented with source and translation pairs to assess in a rage of 0-100. The evaluation set consisted of the sentence pairs annotated by the crowd volunteers and 40 additional repeated segments to account for intra-evaluator reliability. The sentence pairs were randomly ordered in a spreadsheet to avoid potential bias and with no access to any additional information (translation source, crowd annotations, etc).

3.2 Qualitative Feedback

Before looking at evaluation results, this section outlines several impressions of the evaluator, noted during the task in an additional column of the spreadsheet, which have been further developed at write–up. While most have already been discussed elsewhere in the literature, this is yet another opportunity to underscore the relevance of evaluation design for reliable and sustainable results.

	All evaluations		1 evaluation		+ evaluations		2 evaluations	
Trans. source	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
MT1	77.81	23.73	78.93	24.62	77.26	23.32	77.84	23.16
MT2	78.45	23.32	84.83	21.64	75.92	23.52	74.36	23.73
MT3	61.20	29.17	65.35	27.79	58.53	29.79	57.37	29.36
All MT	73.13	26.45	75.97	26.17	71.72	26.49	70.83	46.57
D-MT1	19.89	21.71	23.78	26.87	18.29	19.44	14.39	16.11
D-MT2	21.59	23.91	22.15	23.87	21.35	24.31	21.20	28.31
D-MT3	12.28	18.30	19.93	24.87	7.50	10.72	6.10	7.00
All D-MT	18.20	21.75	19.88	24.69	16.45	20.12	15.50	21.46
Ref	65.99	28.11	75.08	24.59	62.05	28.82	62.25	27.48

Table 2: Overall evaluation results for each translation source (MT systems, damaged outputs and references) reported as quality mean and standard deviation (SD) broken down per number of evaluations collected for each sentence pair (1, 2 or more and only 2).

The effect of (lack of) context. One of the first topics addressed in translator training is text analysis. Numerous scholars have put forward text analysis frameworks to assist translators in this task. To mention an example, the model for translation-oriented text analysis proposed by Nord (1991) establishes that both extratextual factors (author, sender intention, recipient, medium, place and time of production and reception, motive and function) and intratextual factors (subject matter, hierarchy of content and knowledge presuppositions) should be carefully considered as a first step towards drafting a translation proposal. We see, in fact, that a fully developed translation brief involves information that goes beyond providing the surrounding paragraphs or full text where the translated sentence belongs. And it is only after gathering all those details that a translator can make an informed decision on the adequate register, tone, translation strategies, etc. to be used in their target text. The current evaluation set-up presented sentences in isolation (against the recommendation of the latest WMT campaigns, among others). Reportedly, the result of working without context seems to be that the evaluator favours direct translations, which allow to confirm whether all content and nuances of the source are present in the target language, whereas in a contextual evaluation freer translations that move away from the source to display a more natural use of language and better flow of the text would be accepted and even rewarded. This would be possible because the evaluator would be more informed about the importance of the different contents and formal nuances in the sentences. Conversely, without a clear context, these freer translations can appear less accurate and may receive a lower score. This behaviour can potentially promote target language words and structures that are more similar to the source language while discouraging the use of expressions that are natural and specific to the target language. Yet another issue brought by the lack of context seems to be that there are cases where it is not easy to judge the correctness of a translation because of ambiguity or incomplete syntactic structure of the source (that is complemented with a previous or following sentence).

The effect of the source. Aggravated in cases where no context is provided and when non-professionals are involved, the source sentence can become somewhat too referential as to what the best translation would be, and might have an impact on scoring, with the evaluator unfairly supporting close wordings (that are grammatical) while undermining more open possibilities that might be more natural and align better with the tone, register and information flow of a text. This can be of particular interest in language pairs for which language contact –and interference into the minority language– is strong and where the vast majority of speakers of the target minority language are also native speakers of the hegemonic language. A (grammar permitting) word for word translation not displaying any target-language specific expressions and structures could be consistently assessed as excellent translations.

The effect of fluency. Accounting for content transfer in translation can be challenging when sentences use complex structures and when subject-knowledge is needed to fully understand the meaning of the source. In these cases, a fluent translation can be misleading, as extra care is necessary to ensure that all the intended information is present and that no sequences are erroneously interpreted or omitted. This raises the issue of the complexity and thematic typology different evaluator profiles can adequately address.

The precision of the evaluation scale. As a first impression, a 0-100 range seemed very hard to use in the sense that it provided the opportunity to assess quality at a very fine-grained level, while the extent to which mistakes should be penalised seemed greatly subjective. There was a feeling that being consistent with penalisations across the whole evaluation set was hard (see Section 3.3 for agreement results). Admittedly, the evaluator felt more confident with the scale as the number of evaluations performed increased. However, for volunteer work where not a large amount of responses are expected from each individual, a 100-point scale might be a somewhat overwhelming. Note that the professional evaluator wrote a number within an spreadsheet while crowd workers could slide the cursor along a bar, and this might have a impact as well. Additionally, it remained unclear whether the range should be taken as a continuum or a pass/fail threshold should also be considered at 50 points. The annotations collected without the consideration that a score below 50 means that, for example, the translation is unacceptable in a particular situation might differ from those where no such abrupt distinction is made. A similar effect might emerge from scales that use named categories or milestones.

The severity of penalisations. The evaluator reported on the challenge of deciding on a fair penalisation for mistakes. Are 5 points a fair penalisation for an incorrect declension mark? Or should it be 10? 20? 50? Of course, it should depend on the impact it has on the transfer of meaning and on the effect on the form. Even with a context, this is not easy to judge. In fact, impressions noted during the evaluation include a reference to the fact that, not having anchor points to judge the impact of the mistakes and depending on the sentence pair, it would be possible to argue for a score 20 points higher or lower than the one assigned. Moreover, it is not clear whether a penalty should be applied per identified mistake or whether the assigned score should be based on the general impression of the translation quality. To bring a couple of particularly challenging examples, let us consider the cases where incorrect words or short expressions are encountered that do not align with the overall (good) quality of the rest of the sentence; or cases where a fluent translation that does not convey the same meaning of the original but parts of it are completely correct. Unless the texts are for a specific internal use, sentences with any type of error would most probably be deemed unacceptable. They do not fulfil the intended function of the text. As such, a sentence with a meaning or grammar issue would hardly score above a *pass* threshold in a professional setting. However, when presented with a sentence and a 100-point scale, one might one might penalise a mistake with several points but still assign it a good overall*pass*. Sentences might be evaluated in chunks rather than as a unit.

The effect of the perceived translation competence of evaluators. An idea that emerged during assessment was the extent to which the evaluator's perceived language capacity and translation skills could affect the scoring, in other words, whether evaluators project their own translation competence against the translation provided for assessment and judge according to a self-centred threshold. An evaluator might consider any translation that closely approaches to the quality of what they would produce as a good (or not) translation and score consequently; anything better would be highly valued and weaker sequences of that personal threshold penalised. If this risk exists, it might be pertinent to collect information on the self-perceived translation competence (or actual experience) together with linguistic knowledge.

The post-editing effect. Linked to the issue of error severity is the reported temptation to be more lenient towards important mistakes that are easily fixed and to not assign them a heavy penalisation. The incorrect use of a noun or a preposition that changes the meaning of the whole sentence, for example, but can effortlessly be substituted by an adequate noun or preposition without having to tinker with the rest of the sentence elements can feel less damaging. However, in terms of translation quality, the impact of that incorrect element is crucial. If the aim is to collect quality information, ensuring that evaluators can clearly distinguish between translation adequacy and post-editing effort might be relevant.

The quality of the source segments. During the evaluation task, a considerable amount of source sentences was flagged as including grammar or spelling mistakes. While some did not hinder comprehension, others could obscure the correct interpretation of the intended message. An evaluator will not be able to adequately assess the translation quality of a source sentence they cannot understand. For those sentences that could be (adequately) interpreted despite the mistakes, some translations showed no trace of irregularities and were properly resolved. However, at times, the translations do presents mistakes. The question here is whether we want to penalise a translator's inability to overcome issues in the source. The presence of problematic source sentences raises the question of the importance of the configuration of the training or evaluation set we aim to gather. It will probably be a good idea to consider the scale of the evaluation (how much data we can collect) and the specific definition of quality we seek and consciously decide whether we want to include not only correct source sentences but also incorrect ones and even variations and levels of *well-writtenness*.

3.3 Evaluation Results

The evaluator assessed a total of 782 segments. The intraclass correlation coefficient (ICC)[9] calculated with the repeated segments is 0.896 (95% upper bound 0.803 and 95% upper bound 0.9476) which we can interpret as (almost) excellent internal agreement. We visualize these results in an Bland-Altman plot (see Figure 1), where agreement is represented based on the mean difference and by depicting the limits of agreement (Altman and Bland, 1983). If we consider the bias, on average, the second rating of the segments is 2.275 lower, which can be interpreted as a small difference. The data points appear scattered across the graph, indicating the absence of proportional biases or heteroscedasticity. It is also important to note that the great majority of points fall within the limits for the 95% confidence interval, which indicates that the evaluator performs almost equally with the repeated segments. These results can be taken as an indication of consistency in the assessment across the data set. Based on this, we could conclude that the evaluator was able to remain consistent despite the difficulties encountered in the evaluation task and the subjectivity involved in it as described in Section 3.2.

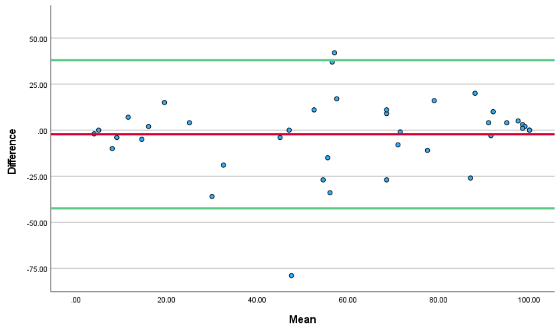

Figure 1: Bland-Altman plot for intra-evaluator agreement for the professional evaluator where the difference between the first and second annotation of the repeated segments is displayed in the Y axis and the average of both annotations is represented in the X axis.

Trans. source	Sent.	Mean	SD	Min.	Max.
MT1	222	78.05	23.07	8	100
MT2	213	79.89	20.94	13	100
MT3	201	58.73	25.26	1	100
All MT	636	72.56	24.94	1	100
D-MT1	32	9.94	8.99	1	30
D-MT2	30	11.07	10.20	1	38
D-MT3	25	8.80	8.71	1	35
All D-MT	87	10.00	9.28	1	38
Ref	59	84.07	18.43	35	100
Total	782	66.47	30.81	1	100

Table 3: Overall evaluation results for each translation source reported as quality mean, standard deviation (SD), minimum score and maximum score according to the annotations of the professional evaluator.

Evaluation results are displayed in Table 3. We can observe that the average score for each of the MT systems is very similar to the scores obtained from the crowd volunteers. These results seem to indicate that overall system quality results would be very similar when evaluated by a translation professional and by (our particular pool of) crowd participants. This is an interesting outcome that might be worth exploring in other evaluation initiatives, as it might be particularly relevant for low-resource scenarios where no funding or professional resources are available for evaluation. It is worth noting that the standard deviations, while large, are slightly smaller than those registered for crowd volunteers and fall within the perceived range of *potential variation* reported by the evaluator (see Section 3.2).

The results for the damaged sentences, however, are up to 10 points lower and all three revolve

[9]Calculated using a two-way mixed model for absolute agreement for a 95% confidence interval).

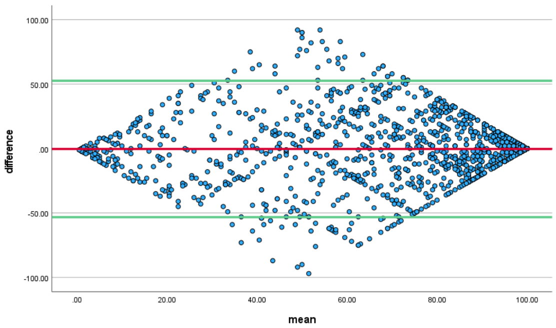

Figure 2: Bland-Altman plot for inter-evaluator agreement between the crowd volunteers and the professional evaluator where the difference between their annotations is displayed in the Y axis and the average of both annotations is represented in the X axis.

around 10 points. Interestingly, even in this case, the ranking for the different damaged sentences follows that of the MT systems that were used to create them. Clearly, the professional translator penalised these cases harsher than volunteers. In contrast, overall, reference translations were assessed almost 20 points higher by the professional evaluator. The differences in these two types of control sentences might indicate that the professional was better equipped to identify the extreme cases and judge them accordingly (see Section 4 for a more thorough analysis of control cases).

In addition to examining quality evaluation results, we also explore the agreement between crowd volunteers and the professional evaluator with respect annotations. Considering all annotations in the data set, the total ICC score is 0.768 (95% lower bound 0.741; 95% upper bound 0.792), which indicates a good agreement (see, for example, Koo and Li (2016) for ICC interpretation). Figure 2 shows a Bland-Altman plot to visualise the overall agreement. The mean difference bias is very close to zero at -0.3125 and we see a random scatter around the mean, mostly within the 95% confidence interval limits. This indicates that the evaluations provided by the two methods observed, that is, a mix of crowd volunteers and a professional evaluator are similar.

If we look more closely, we see that out of the 44 crowd volunteers, when compared with the professional evaluator, three can be assigned an ICC score below 0.5 (poor agreement), 12 an score between 0.5 and 0.75 (moderate agreement), 19 a score between 0.75 and 0.9 (good agreement), and nine a score above 0.9 (excellent agreement).[10] We

Agreement level	ICC	95% lower bound	95% upper bound
Poor	-0.044	-0.729	0.381
Moderate	0.700	0.633	0.754
Good and excellent	0.829	0.804	0.851

Table 4: ICC inter-evaluator agreement between the crowd volunteers and the professional evaluator grouped according to the agreement obtained individually.

calculated the ICC scores for the professional evaluator and the groups of crowd volunteers based on the individual level of agreement obtained. The results show that the ICC agreement with the three evaluators with whom the agreement was poor is actually remarkably poor (-0.044), the ICC agreement with those within the moderate range is rather high within that range (0.7) and the ICC agreement with those within the good and excellent range is very good reaching a 0.829 (see Table 4).

If we consider the individual Bland-Altman plots for each subgroup (Figures 3, 4 and 5), we observe that the bias is moving away from zero as evaluators with a lower agreement are represented in the Figures. The scatter seems to widen when comparing the good and excellent group to the moderate group, but it still shows a random pattern. However, the scatter is clearly not random for the evaluators with a poor ICC agreement.

Table 5 shows the evaluation results in terms of translation quality for the crowd volunteers and the professional translator according to the ICC agreement level groups. We can observe that the average quality assigned by agreeing volunteers is similar, whereas the average of the volunteers which agree poorly with the professional differs in over 12 points. However, what is interesting is that across the groups, the stronger disagreements appear for damaged and reference translations, that is, sentences introduced as control elements to identify evaluator reliability. Differences in MT translations remain the most similar and only appear occasionally as agreement levels decrease. We will consider the performance of these control sentences in more detail in Section 4.

4 Discarding Participants

When evaluators are not asked to work on a minimum number of sentences, it becomes highly challenging to identify outliers because it is not possible to perform consistent comparisons. As an approximation to uncover unreliable participants,

[10] One evaluator only contributed with one evaluation and was therefore not possible to calculate an individual agreement score.

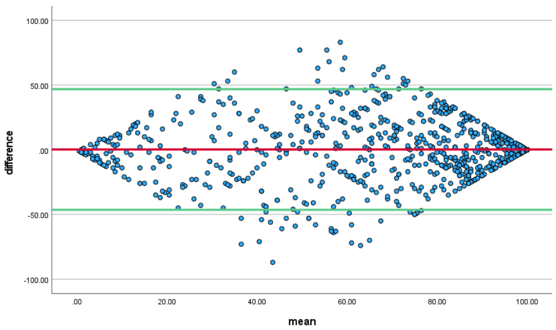

Figure 3: Bland-Altman plot for inter-evaluator agreement between the crowd volunteers and the professional evaluator for which good and excellent ICC agreements were obtained individually, where the difference between their annotations is displayed in the Y axis and the average of both annotations is represented in the X axis.

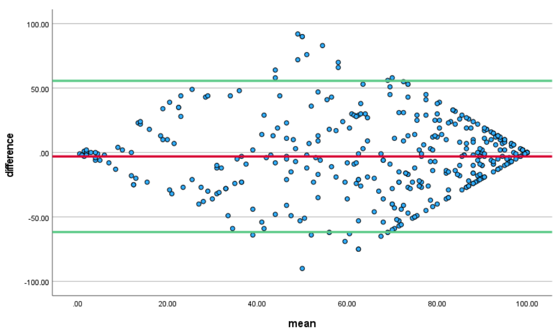

Figure 4: Bland-Altman plot for inter-evaluator agreement between the crowd volunteers and the professional evaluator for which moderate ICC agreements were obtained individually, where the difference between their annotations is displayed in the Y axis and the average of both annotations is represented in the X axis.

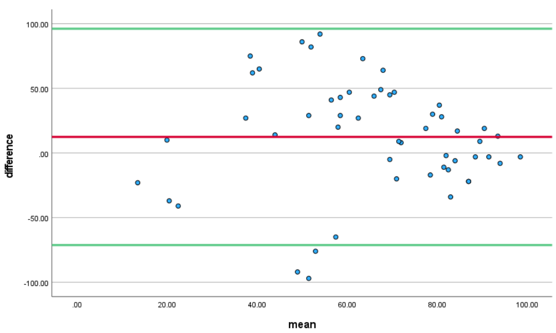

Figure 5: Bland-Altman plot for inter-evaluator agreement between the crowd volunteers and the professional evaluator for which poor ICC agreements were obtained individually, where the difference between their annotations is displayed in the Y axis and the average of both annotations is represented in the X axis.

some works have included in the evaluation set sentence pairs for which the quality is known. It usually involves pairing source segments with artificially damaged translations as examples of bad quality output which should be assessed low, and it can also include source sentences paired with their reference translation as examples of excellent quality pairs. Evaluators who fail to assess them as poor and good cases within a reasonable range are removed from the task and their contributions excluded from the collection. As described in Section 2.2, this is precisely the approach taken by the researchers when running the evaluation task for the data set under study. We looked into these cases to check if these so-called control sentences allow to identify the crowd volunteers which agreed poorly with the professional evaluator.

Out of the 131 damaged sentences included in the data set, 18 were evaluated with a score of 50 or above by 8 crowd evaluators. Out of those 8 evaluators, three have an ICC score ranging between 0.75 and 0.80, that is, a good agreement with the professional evaluator; four have an ICC between 0.61 and 0.72, a moderate agreement; and one of them a low agreement of 0.085, the lowest of all participants. Two out of the three evaluators with a poor ICC agreement with the professional evaluator did not assess any damaged translations. The one who did scored both cases shown with scores above 70 points. This might mean that the inclusion of damaged sentences and its implementation was not particularly accurate in this specific set to filter out deviant evaluators.

To start, not all evaluators assessed damaged sentence pairs. Out of the 44 participants, 36 assessed at least one case. Eight were presented with one damaged sentence pair, 28 were presented with 2 to 11 damaged sentence pairs (Pearson correlation between the total number of evaluations and damaged sentence evaluations is 0.95). Out of them eight failed to pinpoint them as bad quality translations. Only five evaluators with multiple damaged sentence pairs scored them 50 or above more than once. The professional evaluator assigned a low score to all the damaged sentences with scores ranging between 1 and 38.

If we were to discard crowd annotations based on the assessments of damaged sentences, we would be discarding 373 annotations. If we decided to exclude the work of the 8 evaluators who

	poor ICC					moderate ICC					good and excellent ICC				
		crowd		professional			crowd		professional			crowd		professional	
Trans. source	**N**	**Mean**	**SD**	**Mean**	**SD**	**N**	**Mean**	**SD**	**Mean**	**SD**	**N**	**Mean**	**SD**	**Mean**	**SD**
MT1	17	79.47	26.77	71.82	29.64	112	74.62	25.37	78.94	21.18	232	79.07	23.17	77.33	22.86
MT2	15	64.53	31.94	71.93	22.32	108	71.93	25.95	77.93	22.70	228	81.13	22.18	81.62	19.92
MT3	17	66.35	30.91	42.82	24.94	93	62.82	28.57	63.73	25.73	193	59.42	29.42	57.66	24.89
All MT	49	70.35	30.01	61.80	28.99	313	70.74	26.922	74.02	24.02	653	73.98	26.58	73.01	24.66
D-MT1	0	–	–	–	–	21	22.48	30.23	9.33	8.53	31	19.32	18.21	10.55	9.80
D-MT2	3	86.33	11.55	8.67	2.08	10	28.30	22.91	8.00	11.26	37	15.24	18.10	14.68	12.88
D-MT3	0	–	–	–	–	17	19.47	26.25	9.76	10.03	23	8.87	10.60	6.48	5.97
All D-MT	3	86.33	11.55	8.67	2.08	48	22.63	27.10	9.21	9.48	91	15.02	16.885	11.20	10.57
Ref	1	76.00	–	68.00	–	24	58.96	29.84	83.79	16.19	68	71.00	26.98	83.06	19.19
Total	53	71.36	29.17	58.91	30.52	385	63.55	31.28	66.59	31.164	812	67.13	31.676	66.93	30.52

Table 5: Overall evaluation results for each translation source reported as quality mean and standard deviation (SD) for crowed volunteers and the professional evaluator for evaluator groups based on ICC agreements.

were not presented with any damaged translations, we would have to remove another 47 evaluations. This would leave us with a total of 766 evaluations, 64.59% of the total collected. If the damaged translations would have served to accurately identify the outliers (based on the ICC score), we would have only discarded 49 and 47 evaluations, 8.1% of the total evaluations collected.

Let us briefly consider the approach used to create the damaged translations. According to the researchers, these were obtained by translating the source sentences with the three MT systems used to create the remaining data set and by replacing a random sequence of words with a sequence of another source sentence. This approach can result in different types of output, from poor quality target sequences mixed with source language sequences to very good quality target sequences mixed with the source language (note that the quality of the MT systems has been rated within an overall range of 61–78 points). Evaluators trained in translation might be more aware of the importance of the text as a unit and clearly see that such sentences would be unacceptable for the great majority of contexts. Yet this might not be the case for people without translation training. Without a context to consider, it is possible that evaluators do not just penalise the translations for the presence of the source but also feel that they should provide positive points for the sequences with a good quality translation. Depending on the length of the sentences and the proportion of source words, their location within the sentence and the amount of meaning contained in the correct target sequences, it is possible that some evaluators feel that they are being fair by providing a score above 50 to those translations even when they are fully aware of the truncated sequences. Overall, the different results gathered for damaged sentences in this study might indicate that their current design is probably not the most favourable to serve as reliable control sentences.

Together with damaged translations, reference translations were also included in the data set as a control measure. In total, 86 sentence pairs with references were evaluated. Out of the 44 evaluators, 21 were presented with this type of translations: six assessed one case and the remaining 15 assessed from two to 10 cases. Out of the 86 reference sentences, 26 were evaluated with a score of 50 or below by 13 crowd evaluators. Eight of them, which evaluated two or more of such cases, only assigned this score once, whereas the remaining five assigned a low score in multiple occasions. The professional evaluator assessed 11 of the 86 translations with a score of 50 or lower.

Out of those 13 evaluators, three have an ICC score above 0.9, that is, an excellent agreement with the professional evaluator; 11 have an ICC between 0.76 and 0.87, a good agreement; six have an ICC between 0.61 and 0.74, a moderate agreement, and one of them a poor agreement of 0.085, the lowest of all participants. Once again, two out of the three evaluators with a poor ICC correlation with the professional evaluator did not assess any reference translations. The one who did scored the single case presented with a good score of 76, passing the test. This means that the inclusion of reference sentences was not useful in this specific set to filter out unreliable volunteers, on the contrary, by following this test, we would discard good annotations and keep outlier contributions.

As was the case with damaged translations, not all evaluators assessed reference sentences. Out of the 44 participants, 21 assessed at least one case. Six were presented with one reference sentence pair, 15 were presented with two to ten (Pearson correlation between the total number of evaluations and damaged sentence evaluations is 0.93). Out of them 13 failed to pinpoint them as good quality translations. Up to five evaluators with multiple reference sentence pairs scored them 50 or below more than once.

If we were to discard crowd annotations based on the assessments of reference sentences, we would be discarding 632 annotations. If we decided to exclude the work of the 23 evaluators who were not presented with any reference translations, we would have to remove another 374 evaluations. This would leave us with a total of 180 evaluations, 15.18% of the total collected. If the reference translations would have served to accurately identify the outliers (based on the ICC score), we would have only discarded 49 and 47 evaluations, 8.1% of the total evaluations collected.

If we take both control measures into account and combine the performance information of the crowd volunteers, we can account for 37 evaluators out of the 44. If we exclude the work carried out by those who failed any of the tests, we would have to remove the contribution of 18 evaluators, that is, a total of 837 evaluations.

Let us briefly consider the case of reference translations. They were extracted from established sets or other bilingual data published as parallel corpora (Falcão et al., 2024). The quality of reference translations included in test sets has often been questioned and so this was investigated further. If we consider the assessment of the professional evaluator, we see that a score of 50 or below was assigned to seven reference translations out of the 59 presented with scores ranging between 37 and 49. The scores assigned to the remaining references varied from 69 to 100. The range is even wider for crowd volunteers, between 51 and 100. This can be a clear indication that the quality of the reference translations was either not always at a professional level or could not be judged as such out of context. Again, this might mean that carefully choosing high quality references and an evaluation set-up that allows to properly assess their quality is important in order to implement an efficient control measure for non-professional initiatives in particular.

5 Final Remarks

In this work we have explored an opportunistic evaluation initiative that targeted a low-resource language pair (Spanish–Basque) to study the impact of design decisions and the reliability of volunteer participants. A translation professional performed the same evaluation task carried out by volunteers. Next, evaluation results and agreements were compared and the role of control measures that ensure evaluator reliability analysed.

For the analysed set, we can conclude that the overall quality assigned to a MT system might not vary considerably when evaluated by crowd volunteers or a professional evaluator. It remains to be tested if sentence-level accuracy is also as reliable.

In terms of task design, we gathered several issues to consider. Task design is key in providing a working environment that will allow the evaluator to reduce the level of subjectivity and increase consistency. The feedback from a professional evaluator pointed at the benefit of (highly) contextualised sentences, meaningful evaluation categories, manageable complexity and topic specialisation, translation awareness and source sentence quality.

In the same line, identifying outlier contributions seems key to guaranteeing a reliable annotated data set. This being the case, our analysis has demonstrated that while damaged and reference sentences might serve as measures to identify unreliable participants, attention must be paid to creating them. The resulting translations must be unquestionably poor/good so that alternative interpretations are ruled outThen again, it remains to be studied whether participants who properly assess control sentences that are too easily identifiable as poor/good translations will be able to accurately assess regular MT output quality.

All in all, we must not forget that these remarks emerge from the analysis of a single data set and a particular crowd volunteer group. In fact, it would be interesting to study if there are commonalities among the characteristics of the crowd volunteer communities of minoritised languages (participant profiles, level of commitment, level of agreement with professional assessment, for example) and whether these are similar to the crowd participants of hegemonic languages.

Also, this work has explored design issues that are relevant in terms of translation assessment and reliability. However, further research into the real impact of more accurate and cleaner annotations on model training would also be beneficial to determine how rigid (or flexible) an evaluation set-up must be in order to yield useful annotations.

Acknowledgements

This work is supported by the TRAIN (PID2021-123988OB-C31) project funded by MCIN/AEI/ 10.13039/501100011033 and by "ERDF A way of making Europe", DeepR3 (TED2021-130295B-C31) founded by MCIN/AEI/10.13039/501100011033 and European Union NextGeneration EU/PRTR, and the Basque Government (IXA excellence research group IT1570-22).

References

Altman, Douglas G and J Martin Bland. 1983. Measurement in medicine: the analysis of method comparison studies. *Journal of the Royal Statistical Society Series D: The Statistician*, 32(3):307–317.

Aranberri, Nora, Gorka Labaka, Arantza Díaz de Ilarraza, and Kepa Sarasola. 2017. Ebaluatoia: crowd evaluation for english–basque machine translation. *Language Resources and Evaluation*, 51:1053–1084.

Barrault, Loïc, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). ACL.

Bentivogli, Luisa, Marcello Federico, Giovanni Moretti, and Michael Paul. 2011. Getting expert quality from the crowd for machine translation evaluation. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China, September 19-23.

Castilho, Sheila, Maja Popović, and Andy Way. 2020. On context span needed for machine translation evaluation. In Calzolari, Nicoletta, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3735–3742, Marseille, France, May. European Language Resources Association.

Castilho, Sheila. 2021. Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation. In Belz, Anya, Shubham Agarwal, Yvette Graham, Ehud Reiter, and Anastasia Shimorina, editors, *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 34–45, Online, April. Association for Computational Linguistics.

Falcão, Júlia, Kurt Arbela, Nora Aranberri, and Claudia Borg. 2024. Comet for low-resource machine translation evaluation: A case study of english-maltese and english-basque. In *LREC-COLING 2024 - The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Turin, Italy.

Federmann, Christian. 2012. Appraise: an open-source toolkit for manual evaluation of mt output. *Prague Bull. Math. Linguistics*, 98:25–36.

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

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In Pareja-Lora, Antonio, Maria Liakata, and Stefanie Dipper, editors, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria, August. Association for Computational Linguistics.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Kloudová, Věra, Ondřej Bojar, and Martin Popel. 2021. Detecting post-edited references and their effect on human evaluation. In Belz, Anya, Shubham Agarwal, Yvette Graham, Ehud Reiter, and Anastasia Shimorina, editors, *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 114–119, Online, April. Association for Computational Linguistics.

Kocmi, Tom, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 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*, pages 1–42.

Koo, Terry K and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Läubli, Samuel, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium, October-November. Association for Computational Linguistics.

Lison, Pierre and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Nord, Christiane. 1991. *Text Analysis in Translation*. Rodopi, Amsterdam, Atlanta.

Popel, Martin, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Reimers, Nils and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online, November. Association for Computational Linguistics.

Saldías Fuentes, Belén, George Foster, Markus Freitag, and Qijun Tan. 2022. Toward more effective human evaluation for machine translation. In Belz, Anya, Maja Popović, Ehud Reiter, and Anastasia Shimorina, editors, *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 76–89, Dublin, Ireland, May. Association for Computational Linguistics.

Sudoh, Katsuhito, Kosuke Takahashi, and Satoshi Nakamura. 2021. Is this translation error critical?: Classification-based human and automatic machine translation evaluation focusing on critical errors. In Belz, Anya, Shubham Agarwal, Yvette Graham, Ehud Reiter, and Anastasia Shimorina, editors, *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 46–55, Online, April. Association for Computational Linguistics.

Toral, Antonio, Miquel Esplá-Gomis, Filip Klubička, Nikola Ljubešić, Vassilis Papavassiliou, Prokopis Prokopidis, Raphael Rubino, and Andy Way. 2017. Crawl and crowd to bring machine translation to under-resourced languages. *Language resources and evaluation*, 51:1019–1051.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium, October. Association for Computational Linguistics.

Toral, Antonio. 2020. Reassessing claims of human parity and super-human performance in machine translation at WMT 2019. In Martins, André, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 185–194, Lisboa, Portugal, November. European Association for Machine Translation.