COMET for Low-Resource Machine Translation Evaluation: A Case Study of English–Maltese and Spanish–Basque

Júlia Falcão[1,2], **Claudia Borg**[1], **Nora Aranberri**[2], **Kurt Abela**[1]
[1]University of Malta, [2]HiTZ Center, University of the Basque Country (UPV/EHU)
{julia.falcao.22,claudia.borg,kurt.abela}@um.edu.mt,
nora.aranberri@ehu.eus

Abstract

Trainable metrics for machine translation evaluation have been scoring the highest correlations with human judgements in the latest meta-evaluations, outperforming traditional lexical overlap metrics such as BLEU, which is still widely used despite its well-known shortcomings. In this work we look at COMET, a prominent neural evaluation system proposed in 2020, to analyze the extent of its language support restrictions, and to investigate strategies to extend this support to new, under-resourced languages. Our case study focuses on English→Maltese and Spanish→Basque. We run a crowd-based evaluation campaign to collect direct assessments and use the annotated dataset to evaluate COMET-22, further fine-tune it, and train COMET models from scratch for the two language pairs. Our analysis suggests that COMET's performance can be improved with fine-tuning, and that COMET can be highly susceptible to the distribution of scores in the training data, which especially impacts low-resource scenarios.

Keywords: Machine Translation Evaluation, Trainable Metrics, Under-Resourced Languages

1. Introduction

In the early decades of Machine Translation (MT) development, systems were only evaluated by manual methods, but since the early 2000s, automatic evaluation has taken over. BLEU (Papineni et al., 2002), the most popular metric nowadays, measures the lexical overlap between the translation hypothesis and reference translations. Human judgements are still the gold standard (Bojar et al., 2016a), but BLEU is simple and cheap to compute. However, BLEU scores have been shown to correlate poorly with human judgements: BLEU often underestimates systems which humans find better and vice versa, and it also fails to discriminate accurately between high-quality MT systems (Freitag et al., 2022; Mathur et al., 2020a; Kocmi et al., 2021; Callison-Burch et al., 2006; Reiter, 2018).

Other metrics have been proposed to try and mitigate some of BLEU's issues, such as CHRF, which uses character *n*-grams for more flexible matching (Popović, 2015). There are also embedding-based metrics that create soft alignments between hypothesis and reference by consulting resources such as WordNet (e.g. Banerjee and Lavie, 2005; Lo, 2019). However, these are still based only on comparing a hypothesis and reference directly.

In recent years, a new paradigm has emerged: trainable metrics, which are based on neural networks that directly learn to predict human judgements of quality (Mathur et al., 2019; Shimanaka et al., 2018). Their power lies in that they can go beyond the lexical level by generating contextualized representations of the inputs. One of the most prominent has been COMET, a framework for neural MT evaluation models that function as a metric (Rei et al., 2020). COMET models have topped the rankings in recent meta-evaluations, vastly outperforming lexical metrics (Mathur et al., 2020b; Freitag et al., 2021b; Kocmi et al., 2021). Moreover, in the WMT22 metrics shared task BLEU ranked last (Freitag et al., 2022).

Neural metrics like COMET need to be trained on parallel data annotated with human judgements. They are limited in what languages they can support and cover mostly high-resource languages. This limitation has not yet been the subject of systematic analysis, and trainable metrics are most often evaluated on the same languages they were trained on. This poses a large obstacle to their widespread acceptance and use since they compete with lexical overlap metrics like BLEU, which are language-independent and only require tokenization.

The main aim of this paper is to investigate how COMET works for languages outside of its training data and for those unsupported by its underlying encoder. We also analyze the impact of fine-tuning COMET versus training a COMET model from scratch, with Maltese and Basque as the focus of our case study.

Maltese is a Semitic language spoken primarily in Malta, where it is the official language alongside English. Basque is a language isolate and the only pre-Indo-European language in Europe, spoken mainly in the autonomous community of the Basque Country and parts of Navarre in Spain, where it is co-official alongside Spanish, as well as in the French Basque Country (Eberhard and Fennig, 2023). Maltese and Basque are threatened by the presence of major co-official languages in

LREC-COLING 2024, pages 3553–3565
20-25 May, 2024.

their regions, English and Spanish. To ensure that they can continue to thrive in digital environments, machine translation is a key application that facilitates communication and the provision of documentation for native speakers of minority languages. Moreover, it is a tool that can be used to extend resources to include under-resourced languages by automatically translating from English and other higher-resourced languages.

For this case study, we carried out a crowd-based evaluation campaign to collect data in the form of direct assessments of translations from 3 MT systems for each language pair. This dataset is released to public under a CC BY-SA 4.0 license.[1]

2. Related Work

The poor performance of lexical metrics has been a re-occurring topic in WMT metrics shared tasks (Bojar et al., 2016a), and in recent editions (Mathur et al., 2020b; Freitag et al., 2021b, 2022) the best-performing metrics have been trainable systems, most notably COMET (Rei et al., 2020).

COMET is a framework for neural models that function as a metric and it provides both pre-trained models and the means to train custom models.[2] COMET models are built on top of a cross-lingual encoder, usually XLM-RoBERTa (Conneau et al., 2019), and the encoded inputs are passed through a feed-forward network that regresses on human quality scores. This way, models are essentially trained to score translations by predicting how human assessors would judge them, based on contextual embeddings of the inputs.

The default COMET models are trained on Direct Assessment (DA) scores, which are obtained by asking assessors to rate translation hypotheses on a scale of 0–100 (Graham et al., 2013); COMET also supports Multidimensional Quality Metrics (MQM) annotations (Burchardt, 2013), relative rankings (DARR), and Quality Estimation (QE) models (Rei et al., 2021, 2022b, 2023).

Language support is restricted in COMET models due to their underlying cross-lingual encoder: XLM-R, for example, was trained on 100 languages, and cannot reliably encode others; therefore, COMET documentation warns that results for languages outside of this list are unreliable.[3] Moreover, only a subset of these encoder-supported languages are present in the actual training data for each model, so there are languages that are technically supported but not present as source or target in any examples used to train COMET itself. In theory, COMET models should work for all languages supported by the encoder, but this has not yet been the subject of systematic analysis.

COMET models are most often tested on the same language pairs they were trained on, as is the case of its yearly submissions to the WMT metrics shared tasks. One exception is Kocmi et al. (2021), a large-scale meta-evaluation study that covered 232 language pairs supported by COMET but not necessarily present in the training data. This study recommended COMET as the primary metric to use when there is language support, but their results demonstrated that COMET's behaviour is quite unpredictable for unsupported languages.[4] Results seem to vary on a case-by-case basis, in ways that have not been explored in depth. In Mathur et al. (2020b), for example, COMET was evaluated on English→Inuktitut and obtained Pearson correlations of 0.6–0.8, despite Inuktitut not being supported by XLM-R nor included in the training data.

Some researchers have started exploring the potential of extending COMET to evaluate languages outside of its native language support. Sai et al. (2023) evaluated translations from English into 5 Indic languages from 7 MT systems and found that COMET-MQM and COMET-DA obtained the best correlations. They also fine-tuned COMET-MQM on their own MQM scores and reported that the fine-tuned model outperformed the COMET baselines. This study further motivated us to explore such possibilities of improving COMET models for unseen and unsupported language pairs.

3. Preliminary Experiments

Since there were no publicly available parallel datasets with quality assessments for En→Mt or Es→Eu, we leveraged data for other languages first. Our first goal was to examine the impact of fine-tuning COMET and assess the improvements in correlations in comparison to the baseline model.

The current default COMET model, COMET-22 (Rei et al., 2022a)[5], was trained on DA scores from WMT 2017–2020. We looked at the language pairs that were included in the 2021 and 2022 editions and selected 4 pairs with the largest amounts of data, all composed of languages supported by XLM-R: Ukrainian→English (10K samples), Hausa→English (13K samples), English→Hausa (10K samples), and English→Icelandic (10K samples).[6]

[1] https://github.com/MLRS/direct-assessments

[2] https://unbabel.github.io/COMET

[3] https://github.com/Unbabel/COMET#languages-covered

[4] https://github.com/Unbabel/COMET/issues/18

[5] https://huggingface.co/Unbabel/wmt22-comet-da

[6] Links to all the datasets are available in the COMET FAQ.

Model	Uk→En		En→Is		Ha→En		En→Ha	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ
COMET-22	0.017	0.025	**0.423**	**0.589**	0.110	0.159	0.145	0.190
COMET-22-FT	0.099	0.139	**0.488**	**0.673**	0.082	0.114	0.206	0.270
Improvement	**0.082**	**0.114**	0.065	0.084	-0.028	-0.046	0.062	0.081

Table 1: Segment-level Kendall's Tau (τ) and Spearman (ρ) correlation scores for the pre-trained and fine-tuned models on the test set. Scores in red were deemed statistically insignificant (p-values > 0.05).

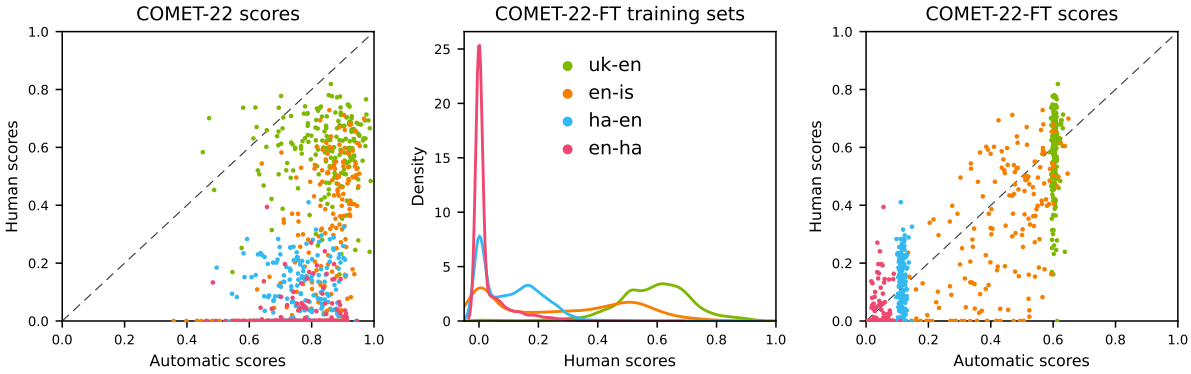

Figure 1: Scores from COMET-22 and COMET-22-FT on the test sets; the training sets are also shown in density plots for comparison with the ranges of scores produced by COMET-22-FT.

In order to fine-tune COMET-22 with these datasets, we first had to rescale the standardized z-scores. As opposed to previous COMET models, which produce unbound scores, COMET-22 returns scores between [0,1] by being trained on z-scores rescaled to this range. Thus, rescaling is necessary as a pre-processing step. The procedure used by its developers[7] is based on min-max scaling[8] the data to $[r_{min}, r_{max}]$, where r_{min} is the mean z-score of all segments with more than 1 annotator where all annotators rated it 0, and r_{max} is the same but where all assessors rated the segments 100. In the subsets of the 4 language pairs we analyzed, all segments were annotated by only one annotator, so we removed this criterion to calculate our r_{min} and r_{max}. After scaling, the scores are clipped to [0,1].

We then randomly picked 10,000 samples for training, 200 for validation and 200 for testing. COMET-22 was originally trained on 900K parallel segments covering 32 language pairs, with datasets ranging from 4K (French→German) to 126K segments (Chinese→English), the median amount being 10K (English→Polish and Lithuanian→English). We further trained it on our datasets, using the hyperparameters recommended in the COMET repository[9], thus generating a new "COMET-22-FT" model for each language pair. We measured the correlation between human scores on the test set and each model's results using Kendall's Tau (τ, Kendall, 1945) and Spearman correlation (ρ, Dodge, 2008).

The results in table 1 show that the correlation scores from COMET-22 are quite low on almost all language pairs, the exception being En→Is which achieves τ = 0.423 and ρ = 0.589. This exception might be due to English and Icelandic being both Germanic languages; we argue that this reduces translation complexity. On the other hand, Ukrainian is a part of the Balto-Slavic family, so it is more distant from English, and Hausa, an Afro-Asiatic language, is unrelated to English.

Overall, Uk→En had the worst scores with COMET-22, which might be attributed to the quality of the data. While the segments for other language pairs were extracted from news articles and most of the COMET-22 training data is also in the news domain, the Uk→En data is made up of real use cases of the Charles Translator for Ukraine (Freitag et al., 2022) and the segments are generally shorter and noisier. Uk→En was the language pair that improved the most with fine-tuning, but correlation scores for COMET-22-FT remained low.

Ha→En was the only pair for which COMET-22-

[7] https://github.com/Unbabel/COMET/issues/138

[8] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

[9] https://github.com/Unbabel/COMET/blob/master/configs/models/regression_model.yaml

FT performed worse than COMET-22, even though fine-tuning brought improvements in the reverse translation direction (En→Ha). Looking at the training logs, we noticed that the Ha→En model performed slightly but steadily better throughout training, but correlations barely increased on the validation set, and given its poor performance in testing, we hypothesize that the model might have overfit on the training data.

We also looked at the distribution of quality scores produced by each model (Figure 1). It shows that after fine-tuning, 3 out of our 4 models only produced scores within a narrow range, which coincide with peaks in concentration of scores in the training data: around 0.0–0.2 for En↔Ha, and 0.6 for Uk→En. The one exception, En→Is, had a more balanced training set, and COMET-22-FT was able to produce a larger variety of scores.

These results brought our attention to the distribution of the data used in training COMET, and how it influences the results that the model is able to produce. It appears that, if the distribution in the training set is unbalanced, the model can severely overestimate or underestimate translations in quality levels that it has not seen in training.

4. Manual Evaluation Campaign

In order to evaluate and train COMET for English→Maltese and Spanish→Basque, we required parallel data annotated with judgements of translation quality. To this end, we ran a crowd-based evaluation campaign to collect human judgements. Due to our limited resources and budget, we designed our campaign for crowd-based, bilingual speakers to participate on a voluntary basis. Similarly to our project, Aranberri et al. (2017) implemented a crowd-based evaluation campaign of English→Basque translations, to investigate whether users found noticeable differences in quality between five MT systems.

We opened the campaign to the general public and shared it with fellow researchers and professionals in related fields. Therefore, as our expected participants would be a mix of experts and non-professional volunteers, it was especially important to be careful with the amount of effort required for the task. We wanted volunteers to be able to participate with minimal effort and to understand the task easily. These concerns were taken into account in the selection of the data, the systems that we evaluated, the choice of our type of evaluation task, and the design of our evaluation tool.

4.1. Data and Systems

For each language pair, we decided to select 400 segments and 3 MT systems. This would yield 1,200 items (source–hypothesis pairs) for evaluation, an amount we deemed viable to achieve through a crowd-based campaign.

	Corpus	Count
En→Mt	FLORES-200 (Goyal et al., 2021)[12]	281
	CrowS-Pairs (Nangia et al., 2020)[13]	49
	EUbookshop[14]	47
	ELITR-ECA[15]	23
Es→Eu	FLORES-200	110
	TED2020 (Reimers and Gurevych, 2020)	60
	Elhuyar[16]	54
	OpenSubtitles (Lison and Tiedemann, 2016)[17]	48
	EhuHac[18]	46
	QED (Abdelali et al., 2014)	40
	WikiMatrix (Schwenk et al., 2019)[19]	30
	NeuLab-TedTalks (Qi et al., 2018)	12

Table 2: Number of segments per original corpus in our datasets.

Our data comprised segments from various parallel corpora (see table 2), most of which were publicly available through OPUS (Tiedemann, 2012).[10] We first tokenized[11] and filtered the segments, to keep those between 10–50 tokens in length. We also used regular expressions to ensure they consisted of well-formed sentences. We proceeded to manually select the segments, carefully avoiding overly specific domains and/or potentially offensive statements against minority groups in society.

As for our choice of MT systems, we selected 3 for each language pair: one proprietary translation engine, one open-source model, and one new model in development at local research groups. The reason for this decision is that we wanted to obtain a wide variety of hypotheses that would then lead to a wider range of scores rather than simply having translations of similar quality. For En→Mt, the chosen systems were Google Translate (GT), NLLB (Team et al., 2022)[20], and UM-

[10] https://opus.nlpl.eu

[11] We used the nltk.tokenize package.

[12] https://github.com/facebookresearch/flores/blob/main/flores200

[13] https://github.com/nyu-mll/crows-pairs

[14] https://op.europa.eu/en/web/general-publications

[15] https://eca.europa.eu

[16] https://elhuyar.eus/en/services/language-services-and-basque-plans/translations-and-language-resources/corpus

[17] https://www.opensubtitles.org

[18] https://ehu.eus/ehg/hac

[19] https://github.com/facebookresearch/LASER/tree/main/tasks/WikiMatrix

[20] We used the 1.3B Dense Transformer variant.

IWSLT (Williams et al., 2023)[21]; for Es→Eu, the chosen systems were Itzuli (the translator on the Basque Government website)[22], NLLB, and UPV-CMBT, a new model in development at HiTZ[23] which is currently not publicly available.

4.2. Evaluation Task

When choosing the type of manual evaluation task to implement, we faced a tradeoff between usability and task complexity. The more information we ask for, the more complex and demanding the task will be for the assessors. We wished to collect scores in a format that would allow us some flexibility for analysis, so we ruled out comparison tasks like pairwise comparison or translation ranking, which only yield relative rankings and do not capture the magnitude of difference in quality. On the other hand, the task could not be as demanding as error analysis frameworks like MQM. Therefore, we decided to collect Direct Assessment (DA) scores, by asking assessors to rate a translation on a scale of 0 to 100 based on how much they agree that "*the candidate translation adequately expresses the meaning of the original text.*"

We chose DA, a method that yields absolute judgements on a continuous scale, because, as argued by Graham et al. (2013), "direct estimates of quality are intrinsically continuous in nature." With absolute scores, we can rank segments and systems as well as estimate the magnitude of difference in quality between them, and we can implement quality control measures to ensure intra-annotator consistency and filter out unreliable assessors. Furthermore, DA scores can be standardized into z-scores, to smooth over differences in the assessors' scoring strategies (Bojar et al., 2016b).

4.3. Platform & Dissemination

We implemented the campaign in a custom version of Appraise (Federmann, 2018)[24], adapted so that anyone could sign in and participate only by creating a username and password. We did not ask users to provide any personally identifiable information, and all instructions were included in the interface in English, Maltese, Spanish and Basque. Once registered, users could evaluate as many segments as they wanted. See the Appendix 11 for more details on the platform.

We ran the campaign for two months, spreading the word through various channels, including university newsletters, mailing lists for linguists and translators, and local social media groups. Table 3 summarizes how many evaluations we received per language pair and the number of participants. Our efforts reached a fair amount of participants, who on average contributed with 24–27 evaluations each. The results reported henceforth are based on the segments that we received evaluations for. These evaluations are being made available to the public to encourage further development and research in the neural evaluation of MT systems.

	En→Mt	Es→Eu
Total evaluations	992	1215
↪ MT outputs	811	996
↪ Damaged outputs	101	133
↪ Reference texts	80	86
Total participants	41	44
Avg. evaluations per user	24	27

Table 3: Human evaluation campaign statistics.

4.4. Quality Control

Among the evaluation tasks, we included two types of control tasks, intended to verify whether assessors were completing the evaluations properly: human references, which should receive higher scores than MT outputs, and damaged MT outputs, intended to be rated lower (Graham et al., 2013; Freitag et al., 2021a). The damaged outputs were created by inserting a random part of a randomly selected reference in the middle of an MT output. This generated sentences that might appear grammatical but did not make semantic sense. To filter our results based on these items, we had to make adaptations to the procedures recommended in the literature, which consisted in comparing an assessor's score for a control item and for its corresponding MT output score; since our participants evaluated varying amounts of items, we could not ensure that they would assess the exact pairs of control items and regular items, so we judged the control items' scores in comparison to all the regular item scores from the same user.

To decide whether a control item was "passed", we dichotomized the DA scale as if users could only rate a translation "good" or "bad", which would correspond to above or below 50. Thus, a damaged output rated below 50 passed, and a reference text rated above 50 also passed. However, different assessors have different scoring strategies, and the absolute threshold of 50 did not account for some cases of assessors with very few evaluations or assessors who, for example, stuck to the top of the scale and only assigned scores between 60–100. Therefore, we also took the standardized z-scores into account; a z-score of 0.0 corresponds

[21] https://huggingface.co/MLRS/translation

[22] https://euskadi.eus/traductor

[23] https://www.hitz.eus

[24] https://github.com/AppraiseDev/Appraise

	En→Mt	Es→Eu
Item occurrences	181	219
↪ Damaged outputs	101	133
↪ Reference items	80	86
Failures	8	11
↪ Damaged outputs	4	2
↪ Reference items	4	9
"Unreliable" participants	5	8
Discarded evaluations	183	361
Remaining evaluations	628	635

Table 4: Numbers related to the quality control filtering procedure.

to the user's mean score, so we considered that a damaged output rated below 0.0 or a reference text rated above 0.0 also passed.

Users who failed one or more control tasks were deemed unreliable, and so we discarded their evaluations. While our strategy might be quite a lenient filter, we believe a deeper analysis would be necessary to improve quality control methods for scenarios in which non-expert participants complete varying amounts of evaluations. For this reason, we release both the filtered and unfiltered datasets, with anonymized usernames, so that other potential studies can look into different filtering strategies. Table 4 summarizes the results of our quality control filtering procedure.

5. System-Level Evaluation

We began our analysis by using COMET-22 to evaluate all of our 1,200 translation hypotheses, obtained from 3 systems for each language pair, in order to rank the systems. For comparison, we also evaluated these hypotheses with 3 lexical metrics: BLEU, chrF (Popović, 2015), and TER (Translation Edit Rate, Snover et al., 2006). We used the implementations from SacreBLEU (Post, 2018)[25], with the default parameters.[26] For all these metrics, the system score is the average of all segment-level scores for each system.

With the human evaluations, we calculated system scores by averaging out the evaluations available for each system. We did this both with the raw DA scores in the 0–100 range and with the standardized z-scores. All the results and their respective rankings are in Table 5.

For En→Mt, the ranking of GT > NLLB > UM-IWSLT is agreed on by humans and almost all metrics. However, looking at the deltas, COMET-22 rates the best and worst systems quite closely, while human raw DA scores are 40% lower. In this case, it seems that the other metrics better capture the degree to which GT is considered better than both NLLB and UM-IWSLT.

As for Es→Eu, human scores and COMET-22 indicate that Itzuli and UPV-CMBT are the best systems and very similar to each other in quality, while the lexical metrics rate NLLB the highest. The BLEU score for NLLB puts it 12 points ahead of both Itzuli and UPV-CMBT, while human participants seem to have found NLLB significantly worse. Nevertheless, like in the En→Mt results, COMET-22 underestimates the magnitude of difference between the 3 systems, rating NLLB only 0.01 less than the others (as opposed to a delta of 18.8/0.53 in human scores/z-scores).

6. Improvement Strategies

The results we collected in our evaluation campaign could also be used to try and improve COMET's performance on our language pairs at training time in two ways: by fine-tuning existing models on this data or training custom models from scratch.

Fine-tuning, as shown in §3, has potential advantages and downsides. COMET-22 is a large model built on top of XLM-R and trained on a parallel dataset of 900K samples. Maltese is not supported by XLM-R, resulting in a vocabulary issue. COMET-22 is much larger than what we can build from scratch with our data. Fine-tuning leverages the existing capabilities of the model and introduces new data for it to be able to handle new languages.

On the other hand, training from scratch—using the framework to create a whole new COMET-DA model—allows us to switch the cross-lingual encoder, and plug in different encoders that were trained mainly on our languages so that the target language embeddings are more reliable and the models might perform better.

Both approaches required a number of additional decisions and pre-processing steps that we describe below, some of which were informed by the findings from our preliminary experiments (§3).

Rescaling Instead of rescaling our z-scores to $[r_{min}, r_{max}]$ and then clipping them to [0,1], we directly rescaled them to [0,1]. We did this because we found we "lost" too many negative scores between $[r_{min}, 0]$ by clipping them to 0, which caused the high concentration of scores around 0.0 that we saw for En↔Ha (shown in fig. 1), and thus severely restricted the distribution of scores in our smaller

[25] https://github.com/mjpost/sacrebleu

[26] BLEU: `nrefs:1|case:mixed|eff:yes|tok:13a|smooth:exp|version:2.3.1`;
chrF: `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1`;
TER: `nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.3.1`.

	System	Automatic metrics				Human evaluation		
		COMET-22 ↑	BLEU ↑	TER ↓	chrF ↑	Raw	Z-scores	Segments
En→Mt	GT	0.7434 #1	43.95 #1	39.49 #1	73.66 #1	82.03 #1	0.593 #1	214
	NLLB	0.6938 #2	24.73 #2	64.47 #3	62.95 #2	65.43 #2	0.116 #2	189
	UM-IWSLT	0.6885 #3	23.82 #3	61.09 #2	59.09 #3	49.11 #3	-0.425 #3	225
Es→Eu	Itzuli	0.8367 #2	15.35 #3	79.20 #3	54.17 #3	82.81 #1	0.439 #1	228
	NLLB	0.8282 #3	27.19 #1	69.36 #1	56.80 #1	63.60 #3	-0.170 #3	192
	UPV-CMBT	0.8371 #1	15.61 #2	78.50 #2	54.36 #2	82.42 #2	0.358 #2	215

Table 5: System-level scores of our 3 chosen systems for each language pair, as assigned by automatic metrics and by participants in our manual evaluation campaign. These scores were calculated by averaging out the segment-level scores. For the human evaluation, we report both raw DA scores and standardized z-scores, as well as the number of evaluations considered for each system.

datasets for En→Mt and Es→Eu. By rescaling directly to [0,1] with min-max scaling, we retain the same distribution along the [0,100] range without the need for clipping.

Stratified Sampling We split our evaluation results into training, validation and test sets so that we could train and evaluate new COMET models. In order to take random samples that were closer to the original distribution of scores, we digitized the z-scores into 10 bins and used stratified sampling to create our test and validation sets, with 100 segments each. The remaining scores made up the training sets (428 for En→Mt and 435 for Es→Eu).

Encoders COMET supports any cross-lingual encoder from HuggingFace Hub that is compatible with BERT, XLM-R, MiniLM or RemBERT architectures. For the models trained from scratch, we plugged in encoders that were specifically developed for the languages we are experimenting with, primarily focusing on each low-resource language. For En→Mt we used mBERTu (Micallef et al., 2022)[27], a version of mBERT (Devlin et al., 2018) that was further fine-tuned with Maltese data. For Es→Eu we chose IXAmBERT (Otegi et al., 2020)[28], a multilingual model for Basque, Spanish and English.

We trained two new models for each language pair: "COMET-22-FT", which is COMET-22 fine-tuned on the training set, and "COMET-DA", which is a new COMET model trained from scratch on the same training set. The hyperparameters used were the same as in §3.

7. Meta-Evaluation Analysis

We evaluated our new models, COMET-22-FT and COMET-DA, in comparison with three lexical metrics and with the baseline COMET-22 model, by computing Kendall's Tau, Spearman and Pearson correlations between the quality scores generated by these metrics and the human scores in the test set. Results can be seen in table 6.

	Metric	τ	ρ	r
En→Mt	BLEU ↑	0.303	0.416	0.399
	chrF ↑	0.368	0.488	0.456
	TER ↓	-0.36	-0.487	-0.440
	COMET-22 ↑	0.292	0.421	0.399
	COMET-DA ↑	0.375	0.527	**0.527**
	COMET-22-FT ↑	**0.391**	**0.542**	0.525
Es→Eu	BLEU ↑	0.021	0.025	0.042
	chrF ↑	0.161	0.236	0.192
	TER ↓	-0.023	-0.022	-0.053
	COMET-22 ↑	0.223	0.326	0.214
	COMET-DA ↑	0.119	0.172	0.169
	COMET-22-FT ↑	**0.245**	**0.354**	**0.242**

Table 6: Kendall's Tau (τ), Spearman (ρ) and Pearson (r) correlation values on the test sets. Values in red are statistically insignificant (p > 0.05).

COMET-22-FT obtained the highest correlations, with an improvement of 0.1 in Kendall's Tau over the base COMET-22 for En→Mt. The delta for Es→Eu was smaller (0.02), but still notable given a training set of only 435 samples. The COMET-DA models performed very differently across language pairs: for En→Mt, it performed better than COMET-22, but for Es→Eu, it obtained low and statistically insignificant correlation scores. Upon closer investigation, we found that all models only produced a narrow range of scores, mainly between 0.6–0.8.

By analyzing the results of the new models COMET-DA and COMET-22-FT, we hypothesize that the distribution of results is influenced by the distribution of the training data. Fig. 2 shows that results are concentrated where most of the scores in the training set are also concentrated. The distribution of training data for En→Mt is slightly more

[27] https://huggingface.co/MLRS/mBERTu

[28] https://huggingface.co/ixa-ehu/ixambert-base-cased

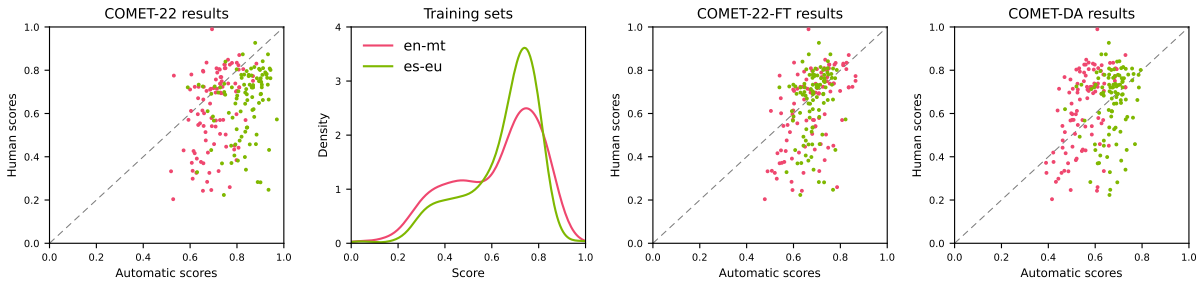

Figure 2: Quality scores generated by COMET-22 out of the box, distribution of the training sets used for our new models, and quality scores from the new models.

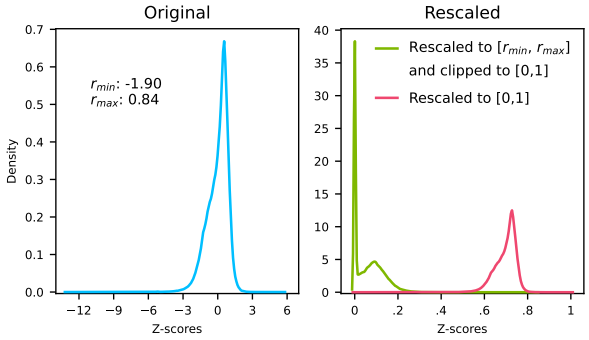

Figure 3: Density plots of the original and rescaled z-scores in the training data for COMET-22.

balanced and the results for this language pair span the range of 0.4–0.9. However, the training set for Es→Eu is highly concentrated around 0.65, and the results are more restricted in the region of 0.6–0.8.

The fact that COMET-22 behaves similarly, producing scores concentrated between 0.6–0.8 despite not having seen any of our data at training time, led us to look closer at its training data as well. We downloaded the datasets from the 2017–2020 WMT metrics tasks and rescaled the z-scores to replicate how they would have been rescaled before training. We also rescaled them the same way we did it for our experiments by min-max scaling directly to [0,1] for comparison. The original and rescaled distributions are shown in Fig. 3.

The z-scores in COMET-22's training set are largely concentrated within [-2.0, 2.0], but outliers bring r_{min} down to -1.9; therefore, all z-scores rescaled to the range of [-1.9, 0.0]—roughly 38% of the training set—are clipped to 0, yielding a very unbalanced distribution of scores. It appears to be the same issue we saw with our training set but on a larger scale. Therefore, if COMET-22 has decent correlations when tested on our data for languages that it technically does not support, it might be out of sheer "luck" as the test data is also mostly within the range of scores the model has seen the most.

In order to test this claim, we made "low-scoring" test sets to evaluate only COMET-22 again, by randomly sampling 100 segments with scores ≤ 0.6 from each training set. The idea is that these MT outputs have been judged as "below average" by human participants, and their z-scores lie outside the range where the training data of COMET-22 is concentrated.

The correlation scores on the low-scoring test set (Table 7) show that the performance of COMET-22 drops significantly in comparison to the regular test set, and the correlations on the low-scoring set are statistically insignificant. Based on this test, we suggest that, in the case of our language pairs, the performance of COMET-22 is unstable and might be heavily influenced by the distribution of scores in its training data.

	Test set	τ	ρ	r
En→Mt	Regular	0.292	0.421	0.399
	Low-scoring	0.099	0.137	0.077
Es→Eu	Regular	0.223	0.326	0.214
	Low-scoring	-0.010	-0.011	0.140

Table 7: COMET-22 correlation scores on the regular and the low-scoring test sets.

8. Conclusion

This paper covered our experiments using COMET to evaluate languages unsupported by its underlying encoder and languages not included in its training data, aiming to evaluate how well it performs out of the box, and how we can improve it by either fine-tuning or by training a new model from scratch. Our findings corroborate those of Sai et al. (2023), who demonstrated that fine-tuning improved the performance of pre-trained COMET models on a set of Indic languages. In our preliminary experiments, by using a training set of 10K samples we obtained improvements of up to 0.08 in Kendall's Tau for Ukrainian→English. We also demonstrate that the scores produced by COMET appeared to be heavily influenced by the distribution of scores in its training data. COMET-22, performed much worse on our low-scoring test set that included only segments scored below 0.6 by humans. Moreover,

the new models, COMET-22-FT and COMET-DA, mostly produced scores between 0.6 and 0.8 after being trained on datasets concentrated in this range of scores. This is an important observation which has a higher impact on low-resource scenarios, when dealing with lower quantities of data to train on and potentially a less diverse set of examples.

More experiments are necessary to confirm these findings, and it might be highly interesting to experiment with training COMET on balanced datasets, even if these are smaller than usual, to see if it can lead to better correlations. Nevertheless, this work is a step towards a better understanding of how pre-trained COMET models work for languages it does not support. We also explored strategies to extend the framework to evaluate low-resource languages, Maltese and Basque. Researchers in the field can further extend this work and adapt our approach for other low-resource languages, so that they will not be left behind as the field of machine translation adopts new evaluation methods.

9. Acknowledgements

This work is based on the first author's master thesis, which was conducted under a scholarship from the Erasmus Mundus European Masters Program in Language and Communication Technologies (EMLCT). We also acknowledge support from LT-Bridge Project (GA 952194) and DFKI for access to the Virtual Laboratory. This work is also based upon work partially supported by the Train (PID2021-123988OB-C31) and DeepR3 (TED2021-130295B-C31) projects funded by the Spanish Ministry of Science and Innovation and the ERDF, and the Basque Government (IXA excellence research group IT1570-22). We are thankful to Kurt Abela and Gorka Labaka for providing us with their MT models for evaluation.

10. Bibliographical References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nora Aranberri, Gorka Labaka, Arantza Díaz De Ilarraza, and Kepa Sarasola. 2017. Ebaluatoia: Crowd evaluation for English—Basque machine translation. *Lang. Resour. Eval.*, 51(4):1053–1084.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016a. Ten years of WMT evaluation campaigns: Lessons learnt. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation–From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 27–34.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016b. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. 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*, London, UK. Aslib.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yadolah Dodge. 2008. *Spearman Rank Correlation Coefficient*, pages 502–505. Springer New York, New York, NY.

Gary F. Simons Eberhard, David M. and Charles D. Fennig. 2023. Ethnologue: Languages of the world (26th edition).

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. 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, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of WMT22 Metrics Shared Task: Stop using BLEU–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

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*, pages 733–774, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *CoRR*, abs/2106.03193.

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*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

M. G. Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

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. *CoRR*, abs/2107.10821.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020a. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020b. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. 2022. Pre-training data quality and quantity for a low-resource language: New corpus and BERT models for Maltese. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 90–101, Hybrid. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Arantxa Otegi, Aitor Agirre, Jon Ander Campos, Aitor Soroa, and Eneko Agirre. 2020. Conversational question answering in low resource scenarios: A dataset and case study for Basque. In *Proceedings of the Twelfth Language Resources and*

Evaluation Conference, pages 436–442, Marseille, France. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Nuno M Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José GC de Souza, and André FT Martins. 2023. Scaling up COMETKIWI: Unbabel-IST 2023 submission for the Quality Estimation Shared Task. *arXiv preprint arXiv:2309.11925*.

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

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022b. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task.

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

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Ananya B. Sai, Vignesh Nagarajan, Tanay Dixit, Raj Dabre, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2023. IndicMT Eval: A dataset to meta-evaluate machine translation metrics for Indian languages.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from wikipedia.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler

Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling human-centered machine translation.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Aiden Williams, Kurt Abela, Rishu Kumar, Martin Bär, Hannah Billinghurst, Kurt Micallef, Ahnaf Mozib Samin, Andrea DeMarco, Lonneke van der Plas, and Claudia Borg. 2023. UM-DFKI Maltese speech translation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 433–441, Toronto, Canada (in-person and online). Association for Computational Linguistics.

11. Appendix A. The Evaluation Platform

Figure 4 shows the registration page. We asked users to choose which language pair they would evaluate, and to select their proficiency level in each language (options were "beginner", "intermediate", "advanced", "fluent" and "native"). All users reported to have at least advanced levels of proficiency in both source and target languages.

Figure 5 shows an example English→Maltese task for evaluation. Participants could evaluate one task like this at a time, and once their score was submitted, they could not go back and change it.

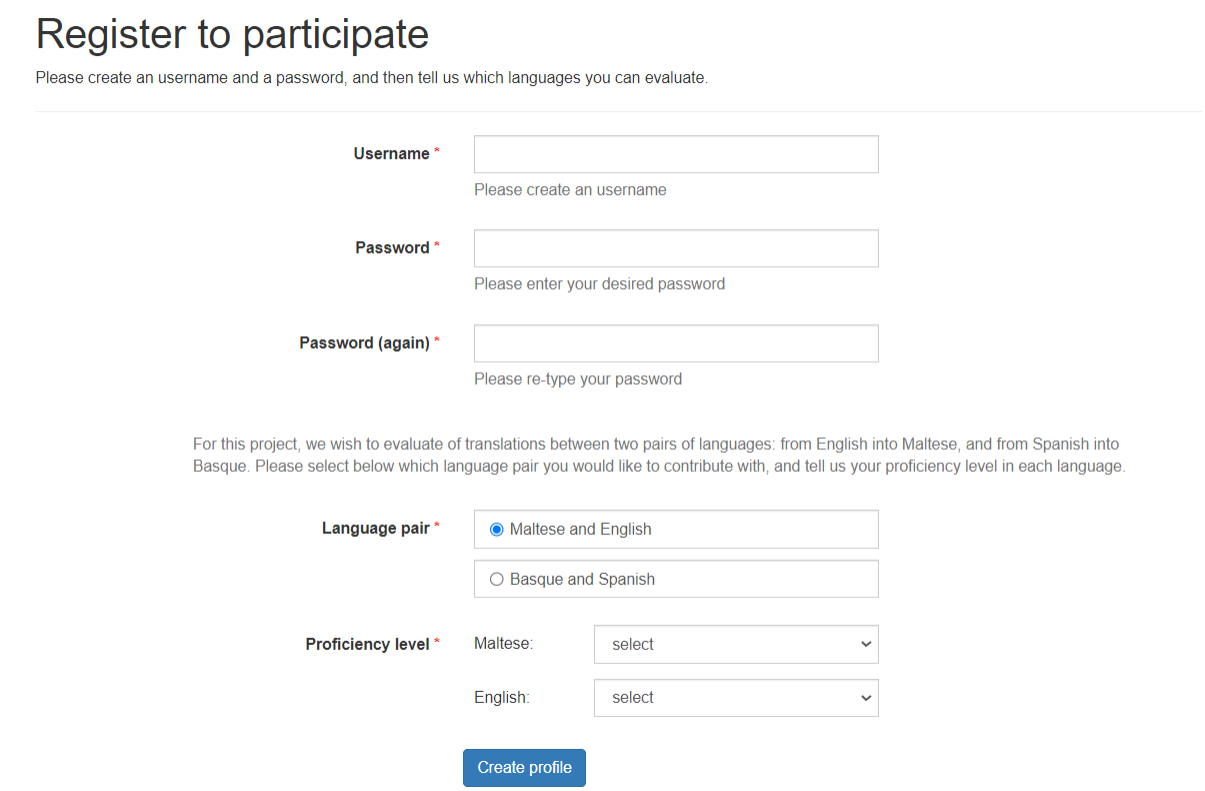

Figure 4: The registration page on Appraise.

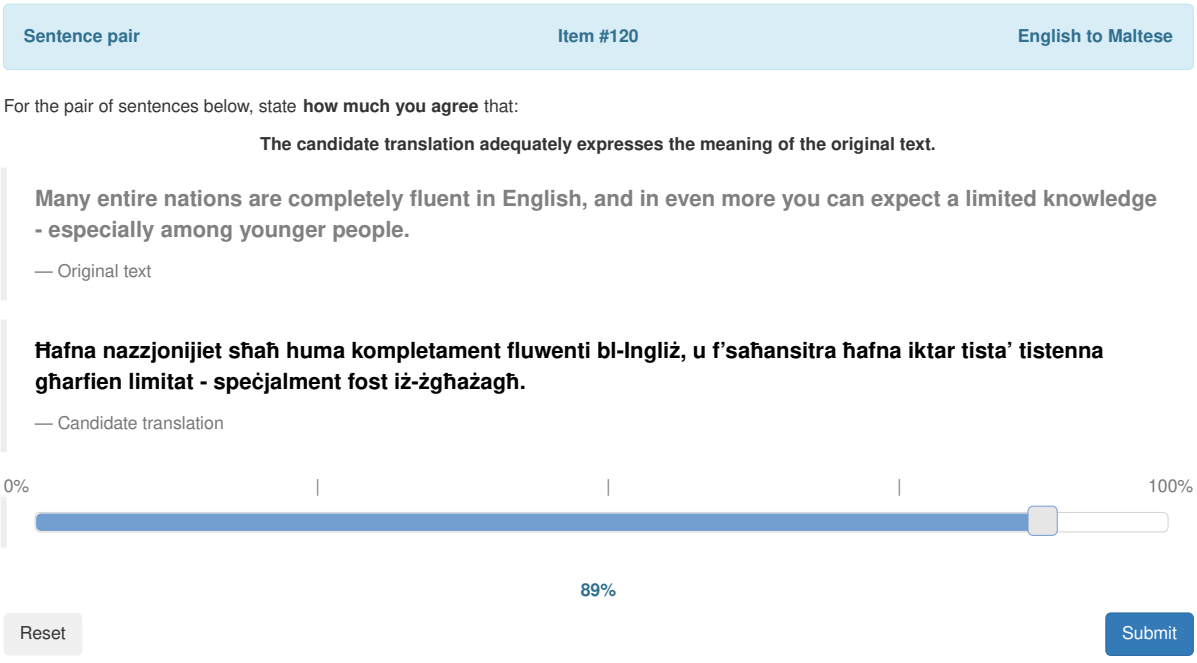

Figure 5: The task page on Appraise.