Global Under-Resourced MEedia Translation (GoURMET)

http://www.gourmet.eu

H2020 Research and Innovation Action
Number: 825299
D5.6 – GoURMET Final progress report on evaluation

Nature	Report	**Work Package**	WP5
Due Date	30/06/2022	**Submission Date**	30/06/2022
Main authors	Sevi Sariisik Tokalac (BBC), Dominic Tinley (BBC)		
Co-authors	Peggy van der Kreeft (DW), Juan Antonio Pérez-Ortiz (UA), Anna Blaziak (BBC)		
Reviewers	Mikel L. Forcada (UA)		
Keywords	evaluation, direct assessment, gap filling, post-editing, benchmarking		
Version Control			
v0.1	**Status**	Draft	15/06/2022
v1.0	**Status**	Final	30/06/2022

Contents

For ease of cross-referencing, the section numbering in this document follows the same structure as deliverable D5.4 Initial Progress Report as far as practicable. Where work was completed in the first half of the project, there are some sections where there is nothing new to report, but this information can be found easily by looking for the equivalent numbered section in D5.4.

Also, where GoURMET languages are listed, they are generally included in order of development with the numbering consistent across this and D5.6 Final Evaluation Report (e.g. details about Macedoniam, language 9, can generally be found in subsection X.9 or X.X.9).

List of Figures

Abstract

This deliverable describes how the GoURMET translation models developed by the research partners have been integrated and evaluated. It follows on from D5.4 Initial Progress Report on Evaluation to describe the final results of the automatic and human evaluation of the translation models.

1 Introduction

1.1 WP5 overview

This document forms part of a series of deliverables that describes how the GoURMET translation models developed by the research partners have been integrated and evaluated.

Work Package 5, coordinated by the British Broadcasting Corporation (BBC) News Labs Multilingual Journalism team comprised five tasks:

T5.1 requirements gathering – see D5.2 Use Cases and Requirements

T5.2 creation of shared interfaces – see D5.3 Initial Integration Report

T5.3 platform integration and deployment – see D5.3 Initial Integration + D5.5 Final Integration

T5.4 media monitoring user evaluation – see D5.4 Initial Evaluation and this report

T5.5 global content creation user evaluation – see D5.4 Initial Evaluation and this report

This document follows on from D5.4 Initial Progress Report on Evaluation to describe the final results of the automated and human evaluation of the translation models.

1.2 Automated Evaluation overview

Automated evaluation assesses the quality of a machine translation system by automatically comparing its output translations to reference translations.

At the time of the interim deliverable D5.4 Initial Progress Report on Evaluation we used *two* standard metrics: BLEU and chrF.

At the time of this deliverable we now use *four* standard metrics: BLEU, spBLEU, chrF and COMET. All four metrics are explained in detail in section 2.1.

1.3 Human Evaluation overview

Human evaluation indicators involve the participation of humans and either collect subjective feedback on the quality of translation or measure human performance in tasks mediated by machine translation.

At the time of the interim deliverable D5.4 Initial Progress Report on Evaluation, we used *two* forms of human evaluation: Gap Filling and Direct Assessment. The methodology for these remains the same as described in detail in D5.4 section 2.2.

1.4 Post-Edit Evaluation overview

At the time of this deliverable, we additionally analyse Post-Edits. The methodology for this additional approach is described in this report in section 2.3 with the results in section 6.1.

2 Evaluation Methodologies

2.1 Automatic Evaluation

Automatic evaluation assesses the quality of a machine translation system by automatically comparing its output translations to reference translations. This enables a quick, cost-effective and reproducible evaluation of a system since, unlike human evaluation, it does not require annotators to directly assess the outputs of the system. However, the ultimate goal of a translation is to fluently and accurately convey the meaning of the source text to users in a language they understand, which is most accurately assessed by human evaluation protocols rather than automated tests. Therefore, automatic evaluation does not replace, rather it complements human evaluation.

Automatic evaluation requires a choice of a test set and an evaluation metric. The test set is a set source sentences with one or more reference translations. Using multiple references can in principle improve the correlation between the evaluation and the value to the user, but in practice obtaining multiple references is expensive and therefore it is not often done in machine translation research. In the GoURMET project in particular we have access to limited amounts of data, therefore we use single reference translations.

The evaluation metric is a function that computes a text similarity score between the generated and reference translations. In this project we use four standard metrics: BLEU, spBLEU, chrF and COMET.

- BLEU (Papineni et al., 2002) is the most common automatic evaluation metric for machine translation reported in the scientific literature. Despite its age and simplicity, BLEU still correlates fairly well with human quality judgements, therefore it is still widely used as the primary, and often unique, evaluation metric in most research papers. It is based on a modified precision computed on word n-grams and corrected by a brevity penalty.

- spBLEU (Goyal et al., 2022) addresses some of the issues of BLEU. BLEU computes the n-gram overlap, which relies on a correct tokenization. This is suboptimal for some low-resource languages, especially for those such as Burmese that don't have word boundaries. Goyal et al. (2022) trained a multilingual SentencePiece tokenizer with the aim to standardise the evaluation accross languages.

- chrF (Popović, 2015) is a metric based on weighted F-scores computed on character n-grams. It has been found to strongly correlate with human quality judgements consistently over different languages and test sets (Ma et al., 2019). Because it is based on characters, chrF is able to give partial scores to word forms which are not in the same morphological form as the reference translation. This is beneficial in the case of morphologically-rich languages (such as Turkish and Kyrgyz) of interest to the project.

- COMET (Rei et al., 2020) is a metric based on a neural model. In contrast with the previous metrics, its computation requires the source sentence. It achieved state-of-the-art correlation with human judgements on the WMT 2019 Metrics shared task (Ma et al., 2019) and has been extensively validated and adopted accross the community (Kocmi et al., 2021).

Different implementations of these metrics exist that can produce slightly different results depending on sentence segmentation, word tokenization and other details. In order to maximise reproducibility and consistency with the scientific literature, we use the implementation of BLEU and chrF

provided by the SacreBLEU tool (Post, 2018a) which has been designed specifically for reproducibility and is widely used. For chrF we use the flag `--chrf-word-order 2`, and for spBLEU `--tokenizer spm`. For COMET we use Unbabel's implementation.

One important feature of ChrF that makes it particularly useful in GoURMET is the ability to give partial scores to word forms which are not in the same morphological form. This is a helpful feature in the case of morphologically-rich languages (such as Kyrgyz) of interest to the project, where word-based metrics such as BLEU would not credit the partial match, for instance, between *şaarda* ('in the city', no possessive marker) and *şaarında* ('in the city', with possessive marker), where the second would be more correct in the phrase *süyüu şaarında* 'the city of love'.

Other evaluation metrics have been proposed in the machine translation literature and they are being evaluated each year in the WMT Automatic Metric shared task (Ma et al., 2019), in some cases obtaining higher correlation with human judgements, but they have drawbacks such as high computational cost, lack of publicly available implementations, limited supported languages, use of machine learning to train the metric (which calls into question their ability to generalise out of their training distributions), and so on.

2.1.1 Evaluation architecture

In order to perform automatic evaluation, we collected a repository of test sets for the GoURMET project language pairs. The following subsections explain the test sets used.

We collected our test sets in the SFTP data repository hosted on the "Valhalla" cluster of the University of Edinburgh. We translated each test set using the Translation Service System Architecture described in deliverable D5.3 section 5 (Secker et al., 2020).

We queried the system using the same API designed for production in order to make sure that our evaluation results are as consistent as possible with the actual use case.[1] Specifically, we sent untokenized source text and received untokenized translations, letting the Translation Service handle tokenization and detokenization internally.

2.1.2 Test sets: m1-m18 translation systems

2.1.2.1 Swahili

The development and test sets were obtained from the GlobalVoices parallel corpus. 4 000 parallel sentences were selected from the concatenation of GlobalVoices-v2015 and GlobalVoices-v2017q3, and randomly split into two halves (with 2 000 sentences each), which were used respectively as development and test corpora. The half reserved to be used as test corpus was further filtered to remove the sentences that could be found in any of the monolingual corpora.

2.1.2.2 Gujarati

The development and test sets are the official sets provided by the WMT19 news shared task (Barrault et al., 2019). The development set contains 1988 sentences. There is a separate test set for each language direction (en–gu and gu–en), so that the source side of each test set is the

[1] except for the English–Tamil system which is still in development and has not been integrated in the Translation Service at the time of this writing.

original text and the target side sentence are the human translations. The en–gu test set contains 998 sentences and the gu–en test set contains 1016 sentences.

2.1.2.3 Turkish

The development and test sets were obtained from the WMT18 news shared task (Bojar et al., 2018). We combined `newtest2016` and `newstest2017` for development, a total of 7,008 sentence pairs, and reserved `newstest2018` for test, a total of 3,000 sentence pairs.

2.1.2.4 Bulgarian

For the test and dev set, we took 4000 sentences from the end of the SETIMES2 corpus (from OPUS). The first 2000 were the test set, and the second 2000 were the test set. The preprocessing was Moses normalisation, tokenisation, truecasing, then BPE with 50k merges learnt separately on each side of the training set.

2.1.2.5 Tamil

The models are currently evaluated using the official WMT20 development and test sets. The development set consists of 1989 sentences. The test sets (separate for each language direction) consist of 1000 sentences for en–gu and 997 sentences for gu–en.

We also created GoURMET development and test set by aligning data from BBC dumps using a modified version of Bitextor (Esplà-Gomis and Forcada, 2010). The models will be tested on these sets at a later date. For document alignment we used an existing MT system to translate all Tamil articles into English and align them based on a TF/IDF score. The alignment was restricted so that only documents originally published within a 30-day time frame of each other are aligned. Segment alignment was done using Bleualign[2], producing a score for each segment pair. For the dev and test set, we took the pairs with a Bleualign score over 0.24 where both source and target sentence contain more than 5 tokens. The sentence pairs were shuffled and split into a dev set of 1916 sentence pairs and a test set of 1917 sentence pairs. Data size is described in Table 1.

Corpus	Sents	en tokens	sr/ta tokens
En-Sr dev	2100	53112	49877
En-Sr test	2100	51933	48762
En-Ta dev	1916	36993	30573
En-Ta test	1917	36940	31180

Table 1: Size of the dev and test sets used for the development and evaluation of the English-Serbian and English-Tamil models. Token counts reported were calculated on raw text, non-tokenised and before BPE segmentation.

[2] https://github.com/rsennrich/Bleualign

2.1.2.6 Serbian

The development and test set for English-Serbian were obtained from the crawled DW corpus. The crawling procedure is described in Deliverable D1.2. From the full En-Sr corpus, we extracted 4200 sentence pairs with a Bicleaner score over 0.8, where both sentences contained more than 10 tokens and the source-to-target length ratio was between 0.8 and 1.1. The Bicleaner model used was an English-Croatian model released with Bicleaner [3]. Half of these sentence pairs formed the dev set, while the other half formed the test set. Data size is described in Table 1.

2.1.2.7 Amharic

The development and test set for English-Amharic were obtained from the GoURMET English-Amharic crawled parallel corpus. There was no provided split for the evaluation sets and therefore we randomly sample sentences. We sample randomly 3,000 unique sentences for each evaluation set.

2.1.2.8 Kyrgyz

A development set and part of the test set were obtained from the GoURMET English–Kyrgyz crawled parallel corpus as follows. First, the crawled corpus was ranked with Bicleaner (Sánchez-Cartagena et al., 2018), whose model was trained on all the publicly parallel corpora for this language pair. Sentence pairs with a score lower than 0.5 were discarded. Then, all the sentences extracted from the news website https://24.kg/ were reserved for the test set. From the remaining sentences, those with a score higher than 0.7, which are very likely to be parallel, were selected to build the development set and the rest was used as training data. The test set was further enlarged with parallel sentences extracted from documents provided by project partner BBC. The number of sentences and words of the test set obtained from each source are depicted in Table 2.

Corpus	Sents	en tokens	ky tokens
GoURMET crawled	144	2 499	1 830
BBC	1 117	19 811	15 749
total	1 261	22 310	17 579

Table 2: Distribution of data among the sources used to build the English–Kyrgyz test set

2.1.3 Test sets: m18-m42 translation systems

For these translation systems we used a private dataset (Table 3) made of processed data dumps in the news domain provided by the user partners. The details on this part can be found in Section 3 of deliverable D1.4.

We also used for evaluation the public dataset FLORES-101 (Goyal et al., 2022), in particular the devtest of 1012 sentences. FLORES-101 is made of sentences from Wikipedia translated into 101 languages by professional translators. Sentences are tagged (in a separate metadata file) with their

[3] https://github.com/bitextor/bicleaner-data/releases

Language	Hausa	Yoruba	Igbo	Urdu	Burmese	Tigrinya	Pashto	Macedonian	Turkish
Sentence pairs	1000	1111	260	1626	1000	600	1350	1000	1633
Origin	BBC	BBC	BBC	BBC	BBC	BBC	BBC	DW	BBC

Table 3: Size of the private datasets for m18-m48

domain (among 10 different domains), which would allow to develop a small-scale domain-level evaluation besides the global one. FLORES-101 was released on June 2021 and all our models are FLORES-101-independent.

Since the private datasets for Igbo and Tigrinya were small, we decided to also evaluate these two languages on publicly available data.

Igbo We used the publicly available test set from Ezeani et al. (2020), which consists of 500 sentences from the news domain.

Tigrinya We extracted 3000 random sentences from the training set, which used several corpora available in OPUS, the Parallel Corpora for Ethiopian languages (Teferra Abate et al., 2018) and the Tigrinya Parallel Corpus from the Travis Foundation (https://github.com/travisfoundation/Tigrinya-Parallel-Corpus). More details on this can be found in Section 2 of deliverable D1.4.

2.1.4 Comparison with Google Translate

We compare our system with the commercial machine translation system provided by Google, at a cost of approximately $20 per million characters (the entire evaluation for all the languages had a total cost of around 200 euros). We submit our test sets to the Google Translate service using their API and we compute the metrics as previously described. The experiments were held in April 2022.

The comparison with Google Translate is essential for our user partners, to help them to calibrate the research models. However, it is not scientifically valid to compare our models with Google. Google's models are not documented and therefore they are not reproducible. We do not know the details of the architecture that they use, when they upgrade or change their translation models from one version to another or the data that they use for training. The most concerning issue is that we cannot exclude that our test sets were contained in the training sets used by Google, since they were extracted from data publicly available on the web. This could lead to artificially inflated scores for the Google system.

The more scientifically rigorous comparisons will always be the results from the annual WMT competition, as WMT use novel test sets produced each year. We have produced Tamil, Gujarati and Hausa test sets for WMT for use in their evaluation campaign and for convincing evidence of our success for the GoURMET project.

2.2 Human Evaluation

See D5.4 section 2.2 for details of the human evaluation framework and methodology.

2.3 Post-Edit Evaluation

One of the aspirations of the evaluation plan was to provide post-editing evaluation as part of the 'gold' standard described in D5.4 section 2.2.2. The Frank prototype was developed to provide the infrastructure for this, with a simple translation comparison and editing window with auto saving functions. The goal was to leave the tool in the hands of intended end users and capture the data for retrospective analysis over a length of time.

Post-editing of machine translation is now very common in the professional translation setting. It is usually understood as an activity where 'a translator compares a source text with a translation produced by an automated process (machine translation or MT) and edits it to make it acceptable for its intended purpose' (Koby, 2012). It is possibly less common in media organisations which repurpose much of their content, rather than translate. However, it is still the case that post-editing machine translation does speed up the creation of new content in another language. For translators there is usually a cut-off point where, if the quality is worse than it, it is more effort to post-edit. Whereas, if it is better, than it is easier to post edit than translate from scratch Zaretskaya et al. (2016). That cut-off point can be measured in post-editing effort.

Post-editing effort can be represented by the number of keystrokes or the number of deletions, insertions and substitutions made. In this research, we have used a quantitative measure of technical effort called *translation edit rate* (TER) Snover et al. (2006), which reflects the number of editing operations necessary to transform the MT output into the final version. Technical effort is more related to cognitive effort than to post-editing time, because post-editing time is strongly dependent on sentence length (Popović et al., 2014). In general, TER is calculated as a number of changed words in the sentence divided by the total number of words.

The post-editing exercise detailed below has been conducted by BBC journalists using BBC World Service content as part of their routine workflows. Due to the editorial concerns around MT accuracy and the resource challenges of production teams, which impacted potential oversight, it was not possible to roll out the Frank prototype to be freely used. The evaluation was therefore conducted with a small subset of trial teams on a limited amount of samples.

The project team selected a subset of language teams on the higher end of the quality scale based on earlier evaluations. This selection was then vetted by another team in view of other ongoing commitments and resource statuses of the language services, and was approved by team leaders.

We settled on three teams: Urdu, Serbian, Urdu and Turkish. The team editors were asked to instruct journalists with proven translation skills to check the prototype for content, and to select around ten (10) articles over the course of ten (10) days.

Approximately half of the articles would comprise original content produced in their language which they regard as worthwhile to promote to other World Service teams in English and the other half would include stories from third languages, translated to their languages via English, which they regard as interesting enough to publish on the respective BBC websites(BBC Serbian, BBC Urdu, BBC Turkish).

The annotators were also asked to fill a survey for each article they have worked on, providing feedback across 12 questions. Since the editorial leaders decided 'duration' would not be an accurate measure, due to journalists having to stop and start the tasks, editing duration was not recorded on Frank. Instead, the survey included questions on how long the editing process took, the kind of edits that were required (e.g. stylistic, factual), and the extent of perceived usefulness of having MT as a starting point for reversions. We describe our results in section 6.1.

3 Interfaces for Human Evaluation

The interfaces for human evaluation were created during the first part of the project, and a full description of the earlier work is available in deliverable D5.4 Initial Progress Report on Evaluation.

3.1 Direct-Assessment Evaluation Tool

See D5.4, section 3.1.

3.2 Gap-Filling Evaluation Tool

See D5.4, section 3.2.

3.3 Open-Source Releases

See D5.4, section 3.3.

4 Results of Data-Driven Evaluation

4.1 Summary of the Results

In Figures 1 (with test sets created from project data - either BBC or DW data) and 2 (with test sets from the publically available Flores data) we report BLEU scores, comparing models trained for the GoURMET project with Google systems, as described in section 2.1.

Google at this time (June 2022) does not have an API which serves Tigrinya translations, even though you can access limited Tigrinya translations via the frontend translate.google.com after a recent exansion of their language coverage by 24 new languages.

We also report other important automatic metrics: in Figure 3, we report spBLEU (*SentencePiece BLEU*, which is robust to tokenisation differences and available for the 101 languages in the FLORES dataset); in Figure 4 chrF score (character-level metric, Popović (2015)); and in Figure 5 the COMET score (Rei et al., 2020) (trained on human evaluations of machine translation and more robust to paraphrases).

We observe that the different scores result in quite consistent rankings of systems for most language pairs and translation directions, hence the three metrics validate each other. This provides evidence that the evaluation methodology is sound, and the rankings reflect a reasonable measure of quality rather than depending on the quirks of a specific metric.

Systems that translate into English mostly obtain higher scores than systems that translate from English. This is expected because there is much more in-domain monolingual data for English than any of the low-resource languages we consider, and monolingual data is most effectively used to improve target-language fluency by means of back translation.

Furthermore, the difference is more pronounced for the BLEU scores than the chrF score. This is expected because English is morphologically simpler than most of the considered languages, which facilitates exact word matching to reference translations, and as discussed in section 2.1, chrF is more robust on morphologically rich languages since it allows for partial word matches.

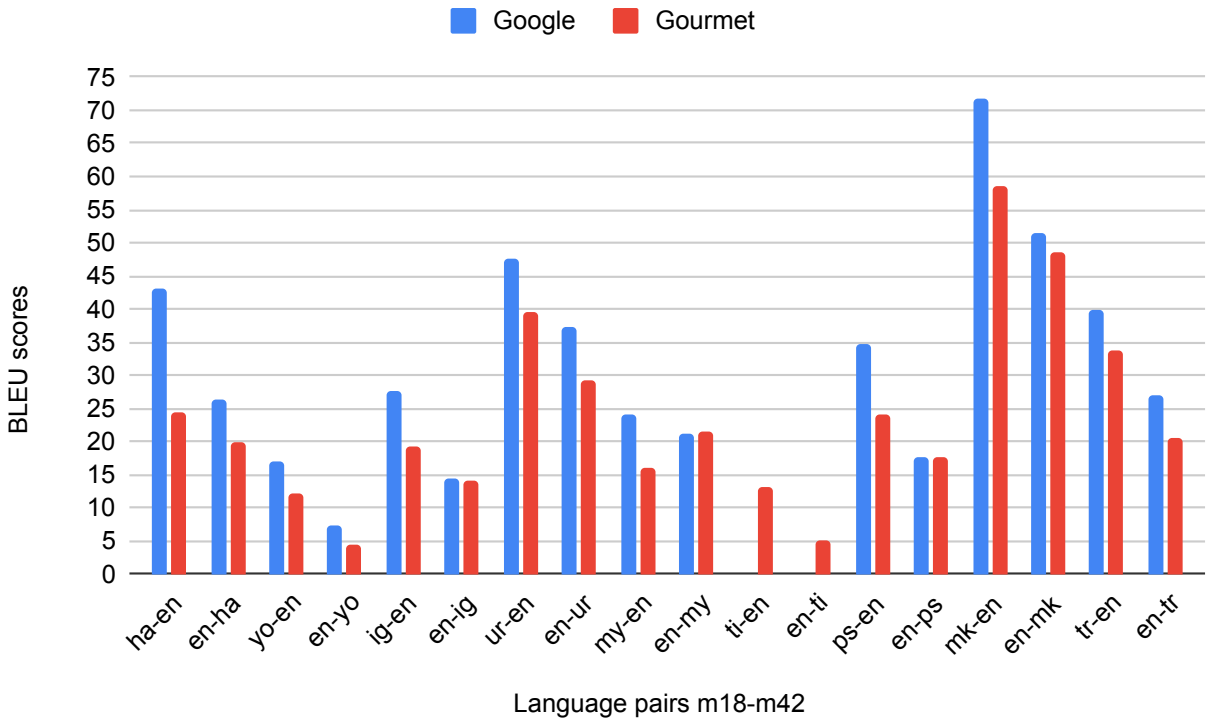


Figure 1: BLEU scores comparing GoURMET m18-m42 translation systems with Google

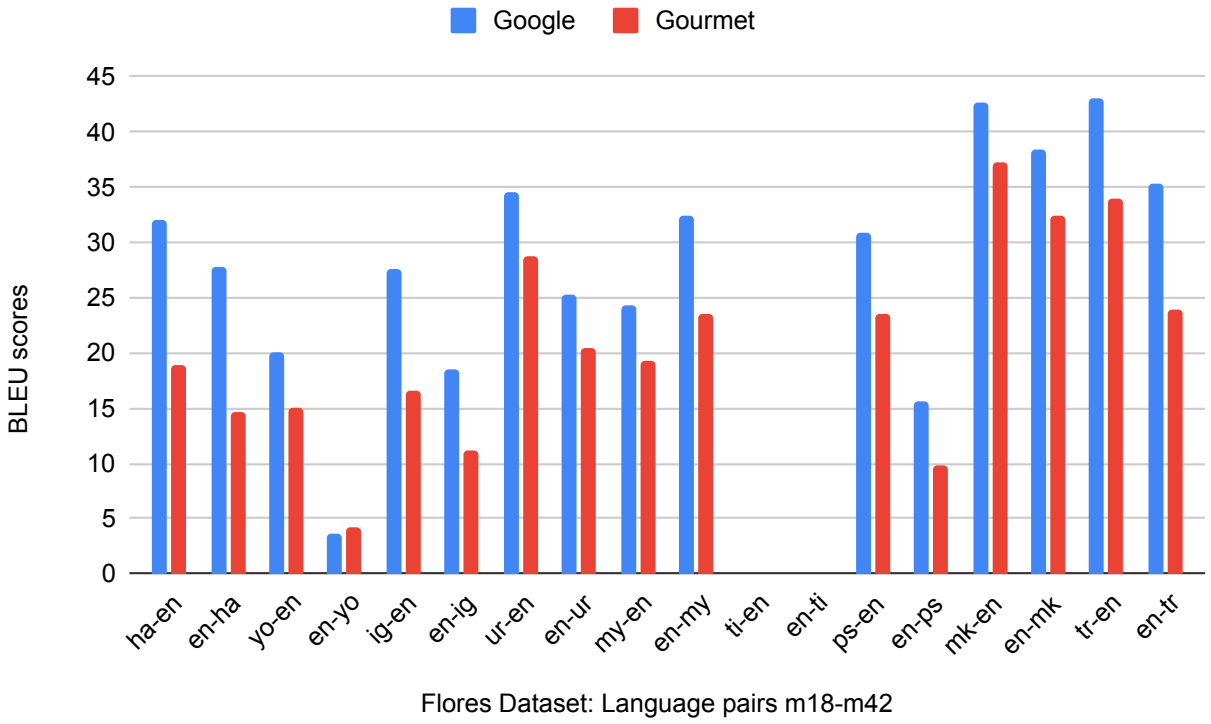


Figure 2: Using the public testset FLORES reporting BLEU scores comparing GoURMET m18-m42 translation systems with Google

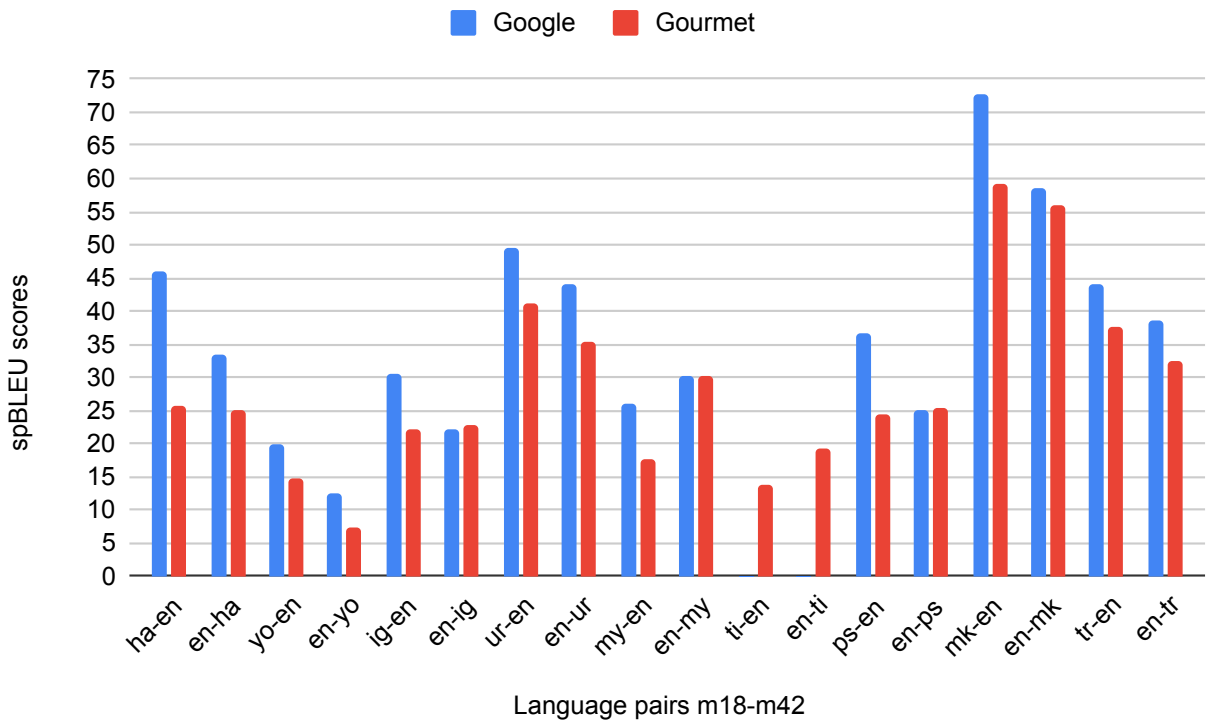


Figure 3: spBLEU scores comparing GoURMET m18-m42 translation systems with Google

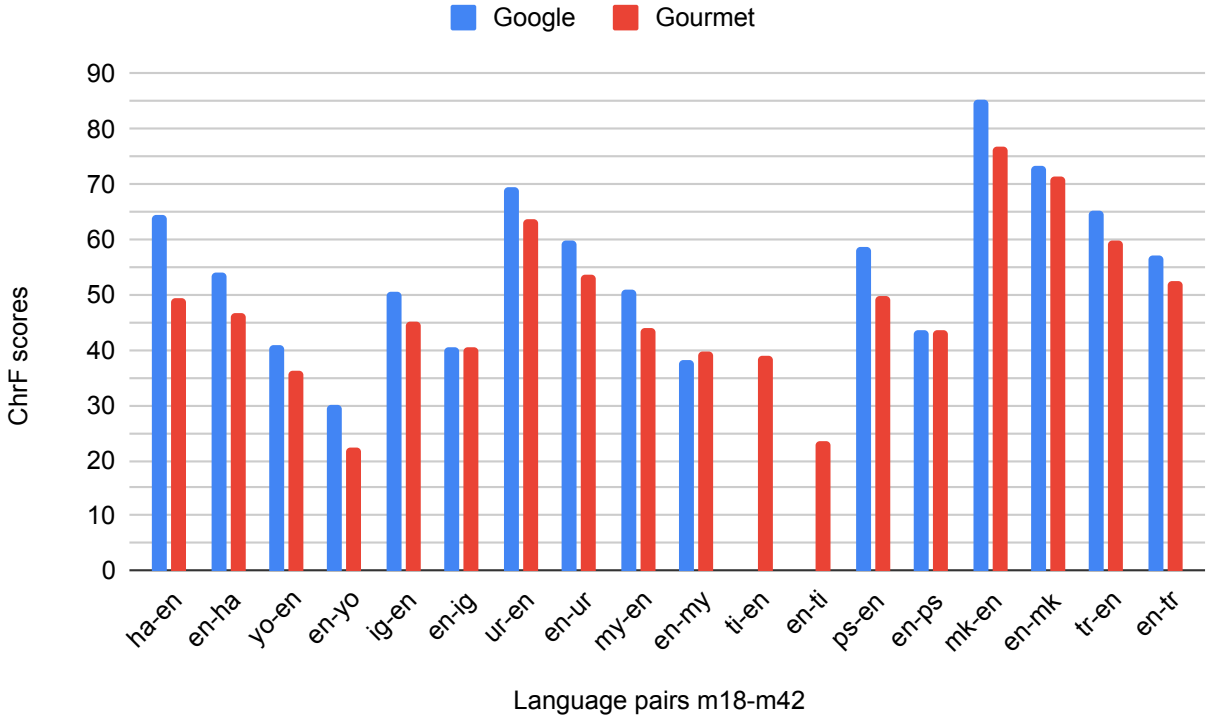


Figure 4: chrF scores comparing GoURMET m18-m42 translation systems with Google

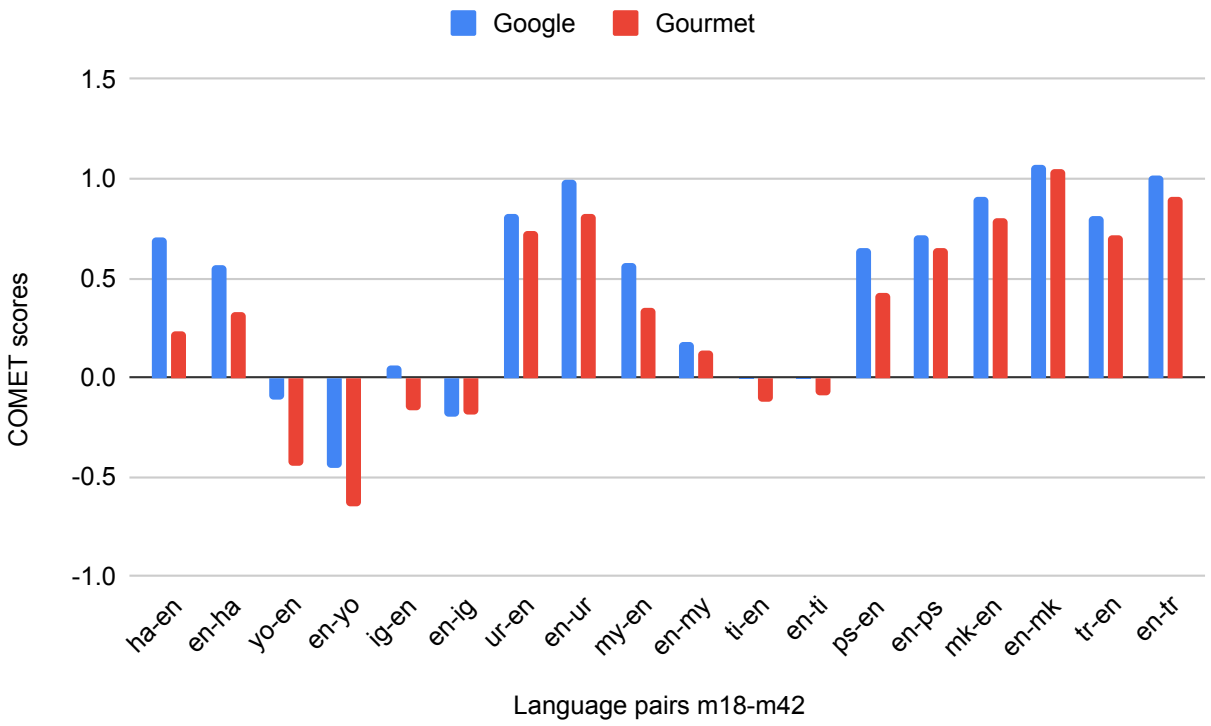


Figure 5: COMET scores comparing GoURMET m18-m42 translation systems with Google

COMET is an interesting metric because it is trained on large amounts of monolingual text, and human judgements of system ranking, and so it is more robust to minor differences in paraphrases. It is worthwhile to note that the COMET scores in Figure 5 largely and closely overlap with the verbal and written testimonials obtained from the human evaluators in this project.

For translation into English, Google Translate obtains equal or higher scores to our systems for most source languages, but for translation from English, our systems are more competitive, surpassing Google Translate for several target languages. These differences might be again attributed to the large amount of English monolingual text that was presumably used by Google to train their systems.

We shall remark that we cannot exclude training-test set contamination for Google Translate, especially for test sets that we scraped from the web and are not part of standard training-test splits (BBC, Deutsche Welle and GoURMET public), hence the scores Google Translate might overestimate its quality.

In Figures 6 and 7 we report BLEU scores for the models from the first 18 months of the project. This is so that we can observe the translation performance of all models delivered in the project in one place. We also address the question of how much have the Google systems themselves improved over the last 24 months. There have been some significant gains for certain language pairs, but most languages have had small gains.

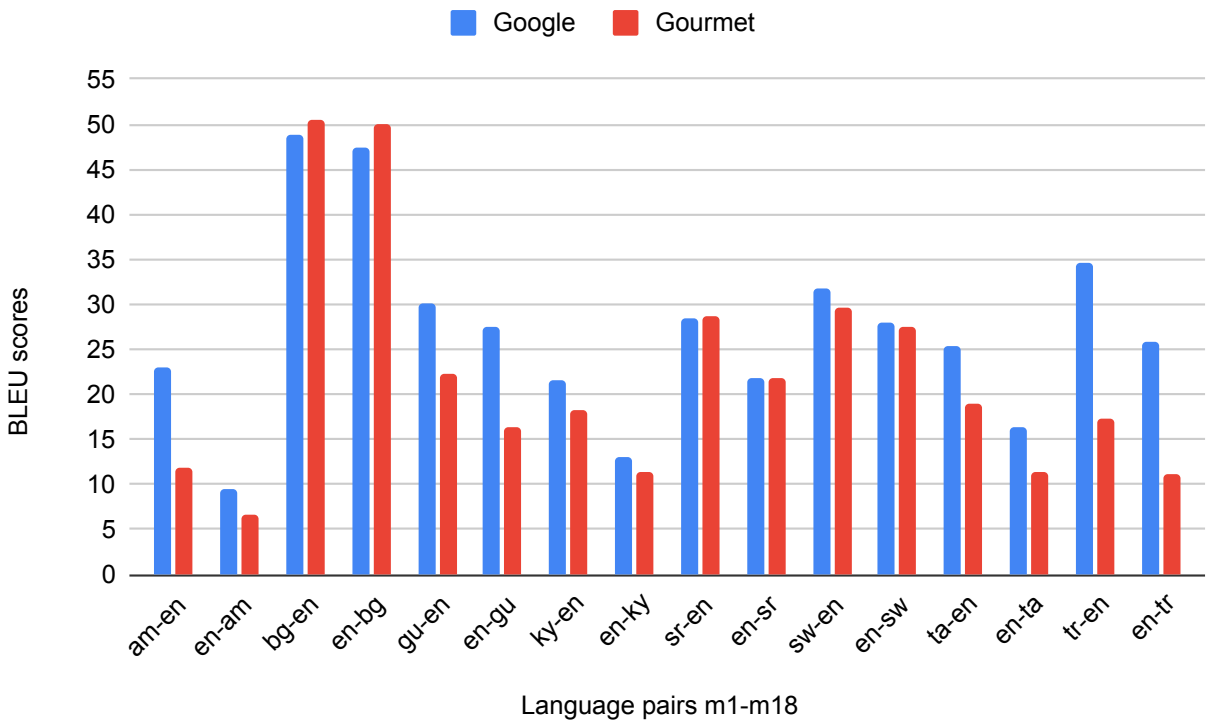


Figure 6: BLEU scores comparing GoURMET m1-m18 translation systems with Google

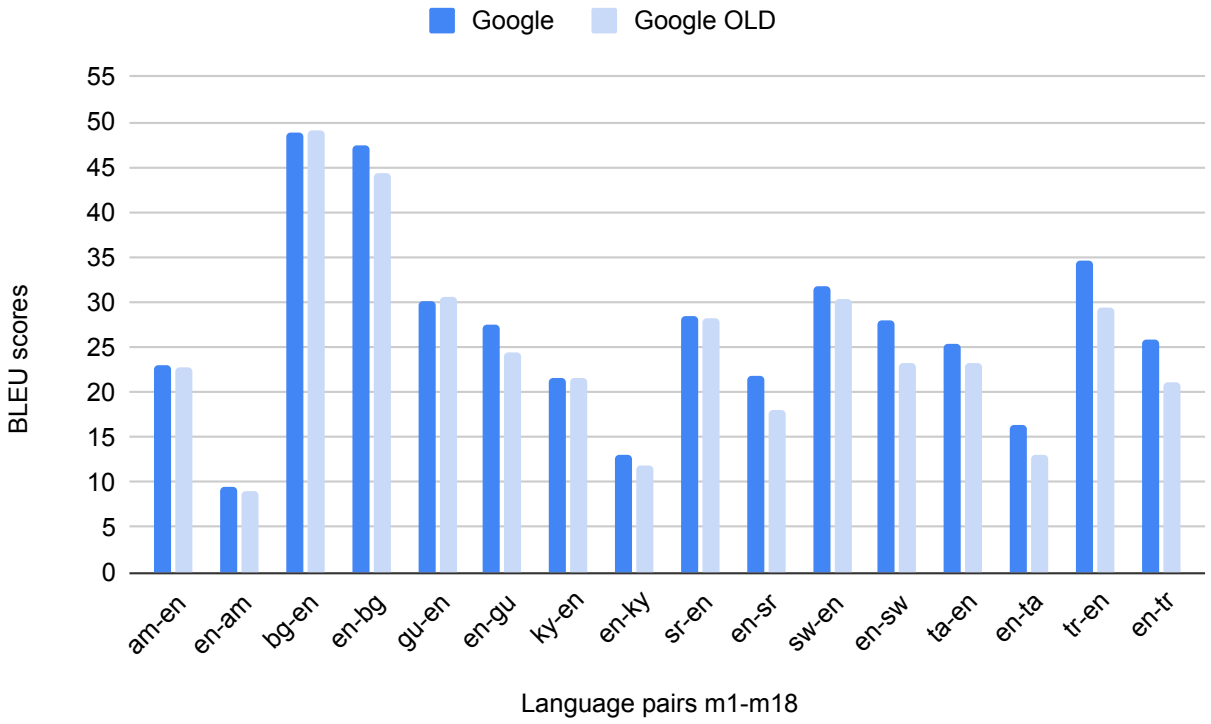


Figure 7: BLEU scores comparing Google systems between m18 and m42

Since the details of the Google Translate platform are unknown, they are irreproducible, and we cannot estimate their computational cost. Any advantage over Google would only ever be temporary as we as a project have made a considerable effort to share our knowledge, our data and our models and do not have mechanisms to continuously improve the models we build.

The advantages of the GoURMET project models are that they can be run in-house for free by anyone, and without the need to share private data with third parties. The benefits from the GoURMET project are not limited to the convenience of our models, but we have also provided tools and data to the community. Furthermore, we have also pushed forward the field of low-resource machine translation by promoting and running low-resource language tasks at the annual WMT competitions.

We have also written a survey paper on the state of low-resource machine translation (Haddow et al., 2022). As part of this survey, we look in more detail at what large industrial research laboratories are doing and how they are successfully training winning systems in the share tasks. But the striking success of large multilingual pre-trained models such as mBART (Liu et al., 2020) and mRASP Pan et al. (2021) still needs further investigation, and massively multilingual models clearly confer advantage to both high- and low-resource pairs (Tran et al., 2021; Yang et al., 2021).

Although these models are successful, further research is necessary to answer questions such as whether the gains are more from the size of models, or from the number of languages the models are trained on, or from the sheer amount of data used. There are also questions about how to handle new languages that are not included in the large pretrained models. This is currently the focus of research in the field.

5 Results of Human Evaluation

There were three main types of Human Evaluation conducted, involving feedback from journalists from the media partners. These were direct assessment (DA) for translations from English into another target language, gap filling (GF) for translations into English, and post-edit evaluation for translations in both directions.

Unlike automated evaluations, which tend to focus on sub-sentence level output, DA and GF evaluations were on sentence level output. In addition to this, post-edits (gold standard) simulated real-life settings where journalists deal with full-length articles and have context to guide them.

5.1 Direct Assessment

As explained in D5.4 sections 2.2.7 and 5.2, all evaluators were asked to rate the quality of the machine translated sentence on a sliding scale from 0% to 100% for two criteria according to the statement "*For the pair of sentences below read the text and state how much you agree that…*", with the criteria being:

Q1 *The black text adequately expresses the meaning of the grey text*

Q2 *The black text is a well written phrase or sentence that is grammatically and idiomatically correct.*

As such, Q1 demonstrates the measure of adequacy of the machine translation and Q2 its fluency.

The next sections show the results of the direct assessment of the translation adequacy of our different NMT systems when translating from English. For convenience, we also present the results for those languages (Bulgarian, Gujarati, Serbian, Swahili, Turkish) already evaluated in deliverable D5.4.

Krippendorff's alpha is used to test interrater reliability and ranges from −1 to 1, with 1 representing total agreement between annotators, and negative values suggesting a systematic disagreement. This interrater reliability measure can also be used to evaluate the degree of agreement between each annotator's score and the corresponding scores assigned to the calibration sentences. Note, however, that in this case the number of samples is so small that one strong difference may drastically lower the overall metric. Calibration sentences were expected to be evaluated as 0/100 or 100/100 in Q1/Q2 scores, respectively.

For all of the languages covered below, evaluations were conducted to silver standard (D5.4 defines silver standard as: Each pair will be rated three or more times (600 responses or more). Three of these languages were selected for gold evaluations, as detailed in section 6.1.

The results from languages covered between m1-18 are also included, since some of the evaluation work was completed after the submission of the D5.4 Interim Report).

5.1.1 Swahili

Status: Silver standard, completed

This section contains the results of the human direct assessment of the English→Swahili system.

The total number of regular sentences in the evaluation dataset is 200. A total of 1044 sentences were evaluated. There are also 10 calibration sentences intended to detect potential misbehaviour.

BBC02	0.29						
BBC03	0.04	0.38					
BBC04	0.09	0.40	0.54				
BBC05	0.09	0.25	0.28	0.36			
DW01	-0.51	-0.06	-0.23	1.00	-0.29		
DW03	-0.13	0.26	0.42	0.60	0.33	1.00	
DW04	0.42	0.45	0.33	1.00	0.21	-0.10	1.00
	BBC01	BBC02	BBC03	BBC04	BBC05	DW01	DW03

Table 4: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Swahili and score Q1. The overall Krippendorff's alpha among all annotators is 0.25.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 4. The pairwise inter-annotator agreements for the score Q2 are shown in Table 5.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 6. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 7.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each evaluator are shown in Table 8. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 9.

BBC02	0.37						
BBC03	0.26	0.34					
BBC04	0.27	0.60	0.48				
BBC05	0.25	0.35	0.42	0.51			
DW01	-0.39	-0.12	-0.11	1.00	-0.26		
DW03	0.07	0.34	0.13	0.46	0.25	1.00	
DW04	0.56	0.57	0.50	1.00	0.23	-0.13	1.00
	BBC01	BBC02	BBC03	BBC04	BBC05	DW01	DW03

Table 5: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Swahili and score Q2. The overall Krippendorff's alpha among all annotators is 0.32.

BBC01	BBC02	BBC03	BBC04	BBC05	DW01	DW03	DW04
0.96	0.98	0.22	0.57	0.67	-0.11	0.16	0.90

Table 6: Calibration agreement computed via Krippendorff's alpha for Swahili and score Q1.

BBC01	BBC02	BBC03	BBC04	BBC05	DW01	DW03	DW04
0.57	0.65	0.39	0.64	0.23	-0.34	0.18	0.77

Table 7: Calibration agreement computed via Krippendorff's alpha for Swahili and score Q2.

BBC01	BBC02	BBC03	BBC04
[73.04, 81.62]	[50.00, 60.94]	[43.50, 50.49]	[35.55, 48.92]
BBC05	DW01	DW03	DW04
[49.15, 57.92]	[9.04, 17.08]	[37.34, 47.16]	[49.40, 61.43]

Table 8: 95% confidence intervals for the true mean of the score Q1 for Swahili.

BBC01	BBC02	BBC03	BBC04
[65.43, 74.78]	[53.26, 63.89]	[47.58, 55.48]	[47.44, 62.96]
BBC05	DW01	DW03	DW04
[52.24, 61.22]	[8.61, 16.43]	[36.94, 47.01]	[50.15, 64.16]

Table 9: 95% confidence intervals for the true mean of the score Q2 for Swahili.

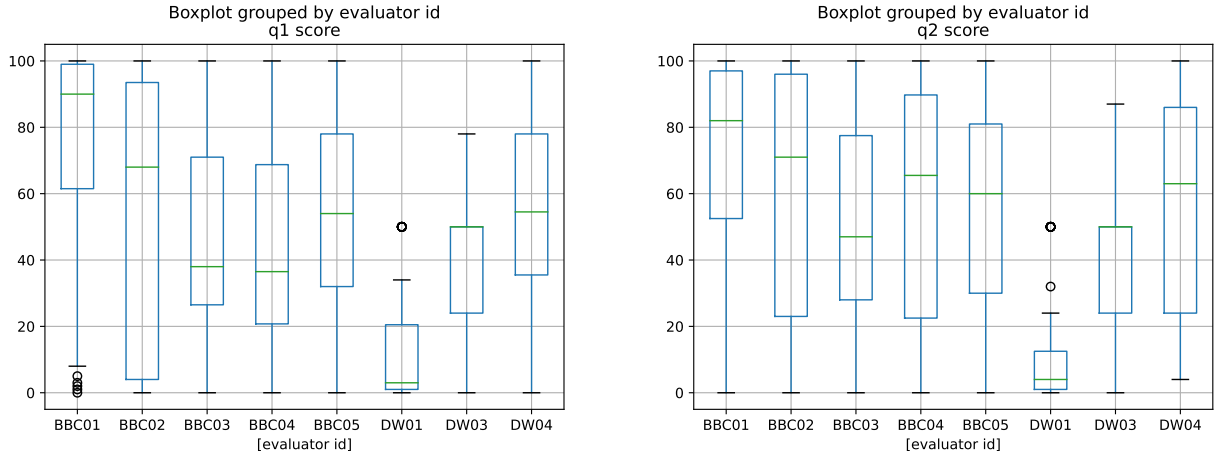


Figure 8: Boxplot of Q1 (left) and Q2 (right) scores for each annotatas or of English→Swahili

Figure 8 shows a boxplot of the adequacy (Q1) and fluency (Q2) scores per annotator.

It is notable that there is little consistency across evaluators with rather large variations across sentences evaluated by each annotator. Overall, fluency is scored marginally higher than adequacy by all annotators except one. However, the majority of evaluators indicate there is degradation in meaning for translations into Swahili.

The main issues that assessors identified with the machine translations were:

Grammar One user commented: 'Most of the sentences were gramatically wrong and distorted the real meaning of what is really intended.'

Meaning One user commented: 'Some second sentences completely changed the meaning of the first, some did not make sense at all.' Another commented that 'some of the machine-generated sentences did not capture the meaning of the sentences that were written by people'.

Sentence length One user commented: 'My impression is that some of the sentences were too long, making it hard to translate.' Another user commented: 'Some sentences are too long to sustain actual meaning when translated.'

5.1.2 Gujarati

Status: Bronze standard, completed

This section contains the results of the human direct assessment of the English→Gujarati system.

The total number of regular sentences in the evaluation dataset is 200. There are also 10 calibration sentences intended to detect potential misbehaviour.

Due to the lack of availability from the BBC World Service Gujarati team due to the Covid-19 situation, despite multiple attempts, only two sets of evaluations had been completed.

BBC02	0.27
	BBC01

Table 10: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Gujarati and score Q1. The overall Krippendorff's alpha among all annotators is 0.27.

BBC02	0.22
	BBC01

Table 11: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Gujarati and score Q2. The overall Krippendorff's alpha among all annotators is 0.22.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 10. The pairwise inter-annotator agreements for the score Q2 are shown in Table 11.

BBC01	BBC02
0.70	0.81

Table 12: Calibration agreement computed via Krippendorff's alpha for Gujarati and score Q1.

BBC01	BBC02
-0.20	0.78

Table 13: Calibration agreement computed via Krippendorff's alpha for Gujarati and score Q2.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 12. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 13.

BBC01	BBC02
[51.76, 59.24]	[43.60, 53.95]

Table 14: 95% confidence intervals for the true mean of the score Q1 for Gujarati.

BBC01	BBC02
[51.92, 59.71]	[44.27, 54.82]

Table 15: 95% confidence intervals for the true mean of the score Q2 for Gujarati.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each evaluator are shown in Table 14. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 15.

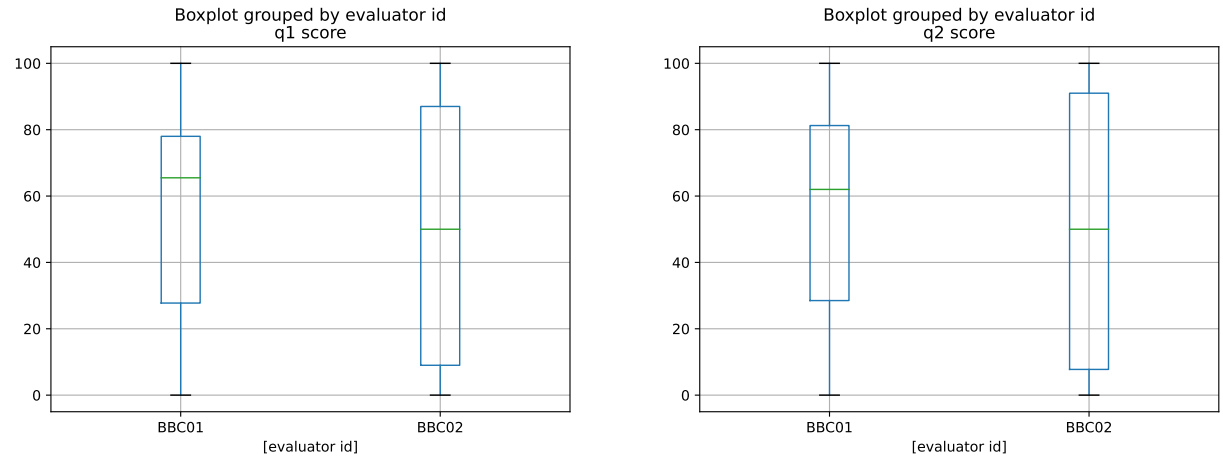


Figure 9: Boxplot of Q1 (left) and Q2 (right) scores for each annotator of English→Gujarati

Figure 9 shows a boxplot of the Q1 and Q2 scores per annotator.

The main issues that assessors identified with the machine translations were:

Proper nouns One user commented that often 'there is a spelling of a name or building that doesnt exist as a word in Gujarati'.

Sentence length One user commented: 'Some of them were quite difficult to decipher, particularly the longer sentences.'

Tense One user commented that issues were mainly 'instances of tense and the odd word being translated incorrectly'.

5.1.3 Turkish v1

Status: Silver standard, completed

This section contains the results of the human direct assessment of the English→Turkish v1 system. See section 5.1.17 for the updated and improved Turkish v2 system.

The total number of regular sentences in the evaluation dataset is 300, a total of 1159 sentences were evaluated. There are also 15 calibration sentences intended to detect potential misbehaviour.

BBC02	0.69				
BBC03	0.72	0.71			
BBC04	0.80	0.74	0.69		
DW01	0.76	0.73	0.67	0.72	
DW04	0.75	0.72	0.77	1.00	0.78
	BBC01	BBC02	BBC03	BBC04	DW01

Table 16: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Turkish v1 and score Q1. The overall Krippendorff's alpha among all annotators is 0.72.

BBC02	0.70				
BBC03	0.68	0.64			
BBC04	0.66	0.69	0.58		
DW01	0.65	0.62	0.55	0.59	
DW04	0.60	0.69	0.77	1.00	0.59
	BBC01	BBC02	BBC03	BBC04	DW01

Table 17: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Turkish v1 and score Q2. The overall Krippendorff's alpha among all annotators is 0.64.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 16. The pairwise inter-annotator agreements for the score Q2 are shown in Table 17.

BBC01	BBC02	BBC03	BBC04	DW01	DW04
0.99	1.00	0.83	1.00	0.94	0.99

Table 18: Calibration agreement computed via Krippendorff's alpha for Turkish v1 and score Q1.

BBC01	BBC02	BBC03	BBC04	DW01	DW04
1.00	0.59	0.51	0.62	0.91	1.00

Table 19: Calibration agreement computed via Krippendorff's alpha for Turkish v1 and score Q2.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 18. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 19.

BBC01	BBC02	BBC03	BBC04	DW01	DW04
[37.79, 48.64]	[36.73, 46.45]	[43.87, 53.64]	[50.93, 62.77]	[34.48, 43.34]	[34.86, 52.75]

Table 20: 95% confidence intervals for the true mean of the score Q1 for Turkish v1.

BBC01	BBC02	BBC03	BBC04	DW01	DW04
[49.04, 60.44]	[54.12, 63.77]	[59.52, 68.57]	[53.97, 65.32]	[49.77, 59.16]	[37.64, 55.28]

Table 21: 95% confidence intervals for the true mean of the score Q2 for Turkish v1.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each evaluator are shown in Table 20. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 21.

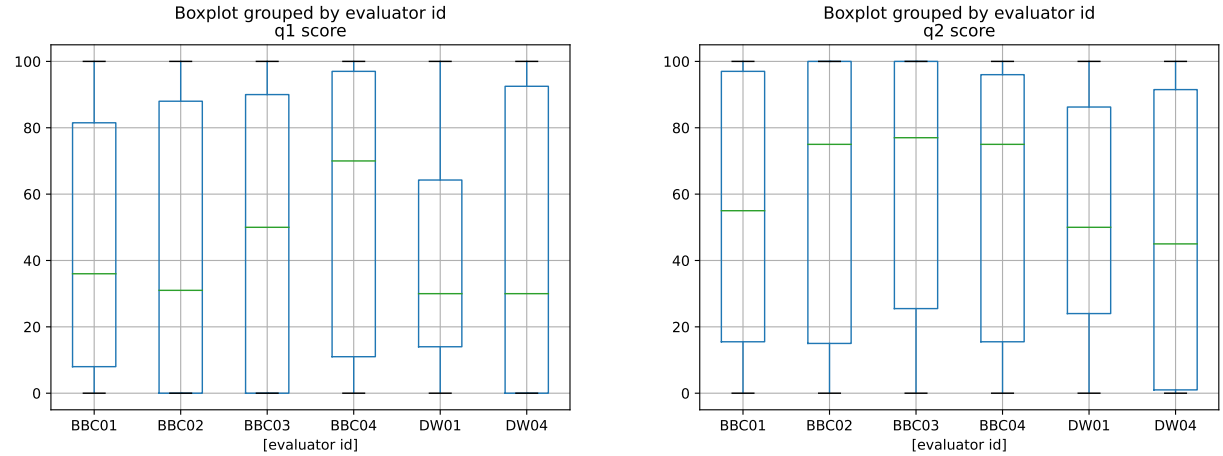


Figure 10: Boxplot of Q1 (left) and Q2 (right) scores for each annotator of English→Turkish v1

Figure 10 shows a boxplot of the Q1 and Q2 scores per annotator.

The range of scores across sentences are among the widest of all languages scored in the evaluations, a pattern repeated consistently across evaluators. This may suggest that the MT output's perceived success rates varied considerably between samples. The scores are more consistent for fluency than for adequacy of the meaning conveyed.

Concerns raised by evaluators about the overall quality of the model prompted additional work on Turkish v2 in the later stages of the project (see section 5.1.17).

5.1.4 Bulgarian

Status: Silver standard, completed

This section contains the results of the human direct assessment of the English→Bulgarian system.

The total number of regular sentences in the evaluation dataset is 200, with 1000 sentences in total annotated.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 22. There are no Q2 scores for this language. There are no

BBC02	0.71			
BBC03	0.69	0.67		
DW01	0.75	0.69	0.66	
DW02	0.74	0.63	0.58	0.74
	BBC01	BBC02	BBC03	DW01

Table 22: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Bulgarian and score Q1. The overall Krippendorff's alpha among all annotators is 0.69.

calibration sentences for the Q1 score for this language. There are no calibration sentences for the Q2 score for this language. Bulgarian was the first language to be evaluated and the calibration practices were further developed after the completion of the work for Bulgarian.

BBC01	BBC02	BBC03	DW01	DW02
[57.87, 68.37]	[56.11, 65.20]	[61.57, 71.70]	[55.85, 66.04]	[46.28, 58.64]

Table 23: 95% confidence intervals for the true mean of the score Q1 for Bulgarian.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each evaluator are shown in Table 23.

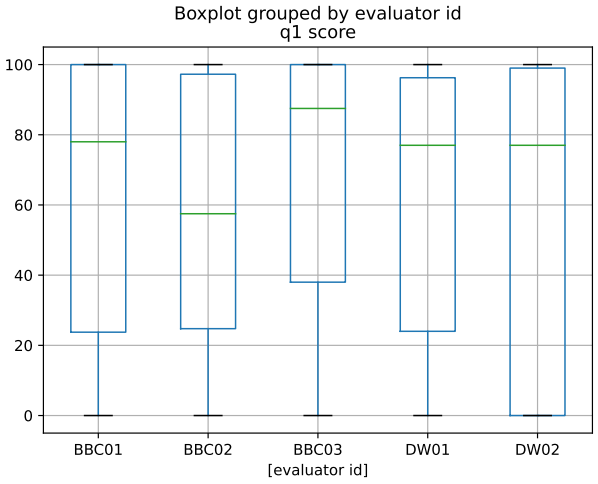


Figure 11: Boxplot of Q1 scores for each annotator of English→Bulgarian

Figure 11 shows a boxplot of the Q1 score per annotator.

Although boxes are tall, hinting at variation in quality, all evaluators indicated that Bulgarian translations achieved top scores in certain sentences.

The main issues that assessors identified with the machine translations were:

Omission One user commented: 'Sometimes the sentence would be almost correct if it wasn't for the lack of a crucial word, such as where the action is happening. But if you have the original text and have some knowledge of the language, that shouldn't be a problem.'

Proper nouns Generated sentences repeatedly failed in translating names, sometimes to confusing effect. In most cases, the name was simply not transliterated into Cyrillic and written out in Latin instead.

5.1.5 Tamil

Status: Silver standard, completed

This section contains the results of the human direct assessment of the English→Tamil system.

The total number of regular sentences in the evaluation dataset is 200, with 900 sentences evaluated by all annotators. There are also 10 calibration sentences intended to detect potential misbehaviour.

BBC02	0.43			
BBC04	0.14	0.58		
BBC06	0.04	0.34	0.10	
BBC07	0.45	0.67	0.51	0.22
	BBC01	BBC02	BBC04	BBC06

Table 24: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Tamil and score Q1. The overall Krippendorff's alpha among all annotators is 0.45.

BBC02	0.30			
BBC04	-0.01	0.47		
BBC06	0.14	0.19	-0.04	
BBC07	0.49	0.46	0.25	0.26
	BBC01	BBC02	BBC04	BBC06

Table 25: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Tamil and score Q2. The overall Krippendorff's alpha among all annotators is 0.34.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 24. The pairwise inter-annotator agreements for the score Q2 are shown in Table 25.

BBC01	BBC02	BBC04	BBC06	BBC07
0.41	0.90	1.00	0.48	0.70

Table 26: Calibration agreement computed via Krippendorff's alpha for Tamil and score Q1.

BBC01	BBC02	BBC04	BBC06	BBC07
-0.82	-0.46	-0.13	-0.66	-0.62

Table 27: Calibration agreement computed via Krippendorff's alpha for Tamil and score Q2.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 26. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 27.

BBC01	BBC02	BBC04	BBC06	BBC07
[25.43, 31.34]	[44.83, 52.61]	[58.76, 69.52]	[44.08, 48.82]	[41.53, 51.03]

Table 28: 95% confidence intervals for the true mean of the score Q1 for Tamil.

BBC01	BBC02	BBC04	BBC06	BBC07
[26.06, 32.00]	[48.44, 56.03]	[65.82, 75.26]	[40.84, 46.20]	[35.49, 43.80]

Table 29: 95% confidence intervals for the true mean of the score Q2 for Tamil.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each evaluator are shown in Table 28. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 29.

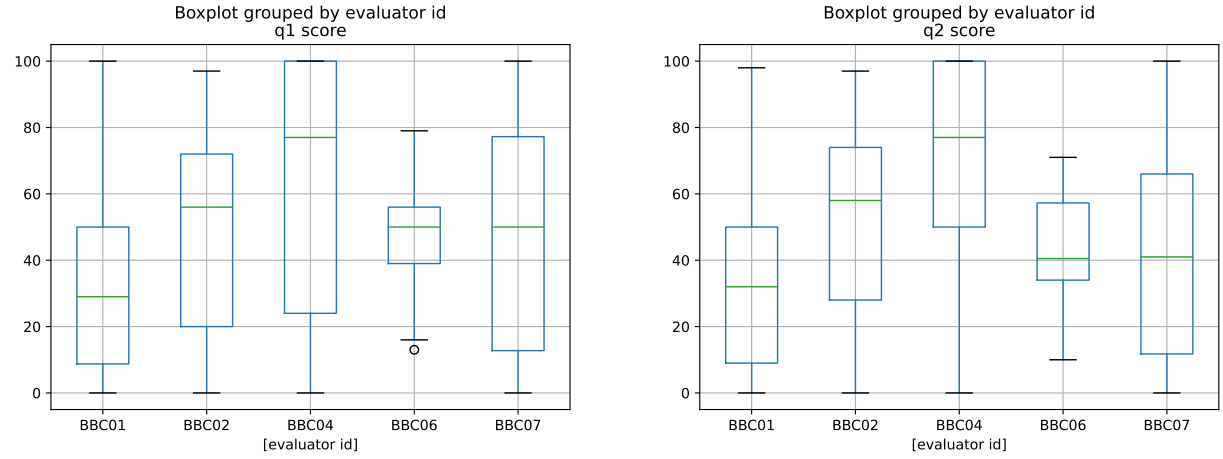


Figure 12: Boxplot of Q1 (left) and Q2 (right) scores for each annotator of English→Tamil

Figure 12 shows a boxplot of the adequecy (Q1) and fluency (Q2) scores per annotator.

It is interesting to note that evaluators have more than 40 points' difference between their median scores, suggesting subjective scoring in evaluations.

The main issues that assessors identified with the machine translations were:

Grammar One user commented: 'Few black sentences adequately convey the meaning of the grey sentences. But they are idiomatically and grammatically wrong.'

Proper nouns One user commented: 'Spellings of nouns like names of persons and places were different in many pairs though they are coherent in terms of meaning, and grammatical and idiomatic correctness. This is a recurring thing throughout the evaluation.'

We also received noteworthy positive feedback:

'Congratulations to the GoURMET team! Your machine translation model looks quite promising and a great alternative to other existing services.'

'My overall impression is that this translation work is simply amazing. To be honest, I thought these sentences were written by human, and I was not aware they were machine generated.'

The range of differing perceptions of usefulness in the comments above are also visible in the range of scores captured by the box charts in Figure 12.

5.1.6 Serbian

Status: Silver standard, completed

This section contains the results of the human direct assessment of the English→Serbian system.

The total number of regular sentences in the evaluation dataset is 200. There are also 10 calibration sentences intended to detect potential misbehaviour.

DW02	0.50			
DW03	0.34	0.37		
DW04	0.17	0.53	0.64	
DW05	0.56	0.45	0.16	0.17
	DW01	DW02	DW03	DW04

Table 30: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Serbian and score Q1. The overall Krippendorff's alpha among all annotators is 0.40.

DW02	0.47			
DW03	0.54	0.69		
DW04	0.13	0.35	0.29	
DW05	0.50	0.52	0.63	0.04
	DW01	DW02	DW03	DW04

Table 31: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Serbian and score Q2. The overall Krippendorff's alpha among all annotators is 0.39.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 30. The pairwise inter-annotator agreements for the score Q2 are shown in Table 31.

DW01	DW02	DW03	DW04	DW05
1.00	1.00	0.95	0.87	1.00

Table 32: Calibration agreement computed via Krippendorff's alpha for Serbian and score Q1.

DW01	DW02	DW03	DW04	DW05
0.99	0.86	0.67	0.45	1.00

Table 33: Calibration agreement computed via Krippendorff's alpha for Serbian and score Q2.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 32. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 33.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each

DW01	DW02	DW03	DW04	DW05
[86.32, 92.28]	[83.43, 93.17]	[71.07, 81.41]	[67.65, 80.45]	[89.17, 96.17]

Table 34: 95% confidence intervals for the true mean of the score Q1 for Serbian.

DW01	DW02	DW03	DW04	DW05
[89.23, 94.83]	[83.60, 93.00]	[85.14, 93.62]	[60.45, 73.59]	[91.80, 97.90]

Table 35: 95% confidence intervals for the true mean of the score Q2 for Serbian.

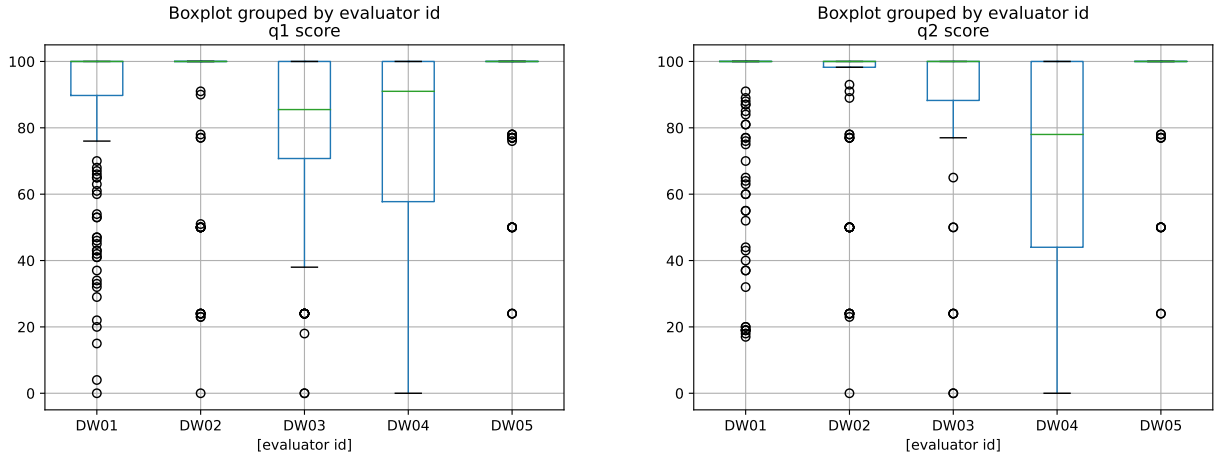


Figure 13: Boxplot of Q1 (left) and Q2 (right) scores for each annotator of English→Serbian

evaluator are shown in Table 34. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 35.

Figure 13 shows a boxplot of the adequacy (Q1) and fluency (Q2) scores per annotator.

Serbian annotators' scores are generally higher than other languages with smaller boxes showing a tighter range of results.

The main issues that assessors identified with the machine translations were:

Gender One user commented: 'In a few cases the gender is translated in a wrong way.' A second user commented 'Most errors are related to the gender or the meaning of the verb.'

Meaning One user commented: 'The machine translation is often shorter and even better for a journalistic sentence. However sometimes it left an important word out and some idioms are inaccurate. Twice or three times everything was completely wrong.'

5.1.7 Amharic

Status: Silver standard, completed

This section contains the results of the human direct assessment of the English→Amharic system.

The total number of regular sentences in the evaluation dataset is 200, with a total of 800 sentences annotated. There are also 10 calibration sentences intended to detect potential misbehaviour.

BBC02	0.27			
BBC03	-0.13	0.04		
BBC04	0.47	0.43	0.26	
BBC05	0.26	0.42	0.38	0.60
	BBC01	BBC02	BBC03	BBC04

Table 36: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Amharic and score Q1. The overall Krippendorff's alpha among all annotators is 0.41.

BBC02	0.07			
BBC03	-0.14	0.06		
BBC04	0.29	0.24	0.29	
BBC05	0.28	0.41	0.37	0.55
	BBC01	BBC02	BBC03	BBC04

Table 37: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Amharic and score Q2. The overall Krippendorff's alpha among all annotators is 0.36.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 36. The pairwise inter-annotator agreements for the score Q2 are shown in Table 37.

BBC01	BBC02	BBC03	BBC04	BBC05
0.41	0.83	0.69	0.40	0.82

Table 38: Calibration agreement computed via Krippendorff's alpha for Amharic and score Q1.

BBC01	BBC02	BBC03	BBC04	BBC05
-0.40	-0.63	-0.35	-0.51	-0.44

Table 39: Calibration agreement computed via Krippendorff's alpha for Amharic and score Q2.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 38. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 39.

BBC01	BBC02	BBC03	BBC04	BBC05
[41.90, 53.90]	[62.43, 67.99]	[77.32, 83.95]	[50.32, 60.51]	[62.76, 72.03]

Table 40: 95% confidence intervals for the true mean of the score Q1 for Amharic.

BBC01	BBC02	BBC03	BBC04	BBC05
[42.73, 56.87]	[71.70, 76.12]	[76.88, 83.46]	[53.94, 64.25]	[67.01, 75.36]

Table 41: 95% confidence intervals for the true mean of the score Q2 for Amharic.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each

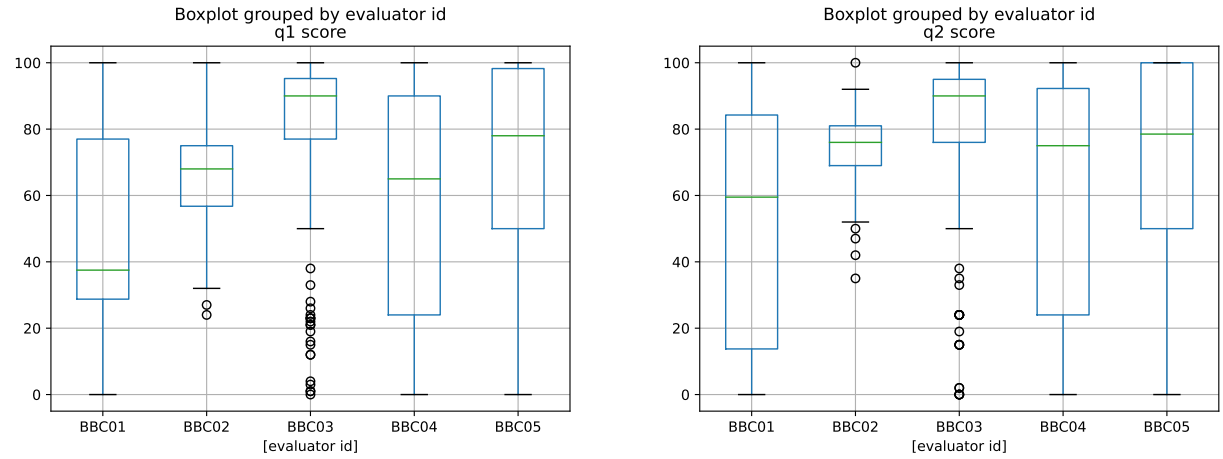


Figure 14: Boxplot of Q1 (left) and Q2 (right) scores for each annotator of English→Amharic

evaluator are shown in Table 40. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 41.

Figure 14 shows a boxplot of the Q1 and Q2 scores per annotator.

The overall scores are on the higher end of the scale, despite the existence of numerous outliers. It is worth bearing in mind that Amharic was initially returning errors when evaluation was first attempted and it was thus one of the languages that had seen the model revised (as noted in D5.5 section 8.4).

The main issues that assessors identified with the machine translations were:

Alphabets One user commended: 'In Amharic there are few alphabets with similar sound but each are used in different words for different meaning' and another that 'Amharic letters [Ethiopic Syllable Glottal Aa] and [Ethiopic Syllable Glottal A] are used interchangeably. The sound is the same but [Ethiopic Syllable Glottal Aa] is used in Tigrinya more often.'

Punctuation One user commented: 'If this project is all about assessing the translation level, the missing punctuation may not be a big deal for now. Still, some important punctuation marks bring about massive meaning change, just like a comma in an English text.'

Sentence length Several users commented that shorter sentences were translated better than longer sentences.

5.1.8 Kyrgyz

Status: Silver standard, completed

This section contains the results of the human direct assessment of the English→Kyrgyz system.

The total number of regular sentences in the evaluation dataset is 200. There are also 10 calibration sentences intended to detect potential misbehaviour.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 42. The pairwise inter-annotator agreements for the score Q2 are shown in Table 43.

BBC04	-0.35		
BBC05	0.35	-0.21	
DW03	0.23	0.08	0.22
	BBC02	BBC04	BBC05

Table 42: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Kyrgyz and score Q1. The overall Krippendorff's alpha among all annotators is 0.14.

BBC04	0.06		
BBC05	0.44	-0.03	
DW03	0.14	0.41	0.04
	BBC02	BBC04	BBC05

Table 43: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Kyrgyz and score Q2. The overall Krippendorff's alpha among all annotators is 0.22.

BBC02	BBC04	BBC05	DW03
0.67	0.47	0.99	0.89

Table 44: Calibration agreement computed via Krippendorff's alpha for Kyrgyz and score Q1.

BBC02	BBC04	BBC05	DW03
0.19	0.39	0.41	0.38

Table 45: Calibration agreement computed via Krippendorff's alpha for Kyrgyz and score Q2.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 44. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 45.

BBC02	BBC04	BBC05	DW03
[41.31, 48.36]	[82.82, 87.51]	[34.27, 44.69]	[58.43, 65.90]

Table 46: 95% confidence intervals for the true mean of the score Q1 for Kyrgyz.

BBC02	BBC04	BBC05	DW03
[43.76, 52.10]	[74.28, 80.50]	[36.00, 47.26]	[74.34, 79.96]

Table 47: 95% confidence intervals for the true mean of the score Q2 for Kyrgyz.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each evaluator are shown in Table 46. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 47.

Figure 15 shows a boxplot of the Q1 and Q2 scores per annotator.

It is worth noting that Kyrgyz is one of the languages with the most inconsistency among scores by different evaluators. This pattern is also on view in the Krippendorff's alpha scores which are the lowest among all languages.

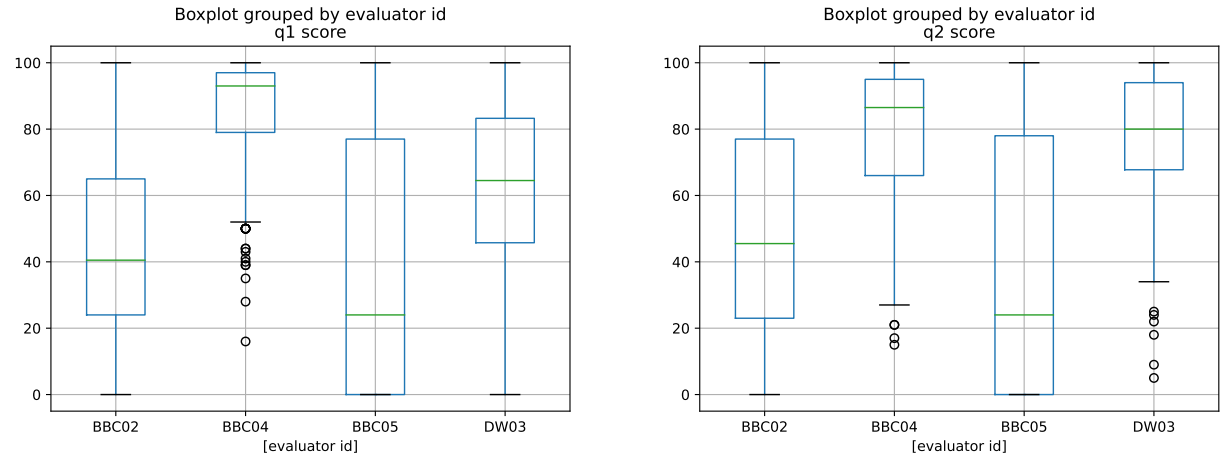


Figure 15: Boxplot of Q1 (left) and Q2 (right) scores for each annotator of English→Kyrgyz

Several evaluators noted it felt like the sentences were hit and miss, with good quality translations mixed in with random translations. Two evaluators said their experience was overall better than what they previously had with Google.

The main issue that assessors identified with the machine translations was:

Sentence length One user commented: 'The machine tries but in some cases, it is obvious that it is not the proper manner of speaking.' They conclude: 'But it is a good effort. Well done, Machine.'

5.1.9 Macedonian

Status: Bronze standard, completed

This section contains the results of the human direct assessment of the English→Macedonian system.

The total number of regular sentences in the evaluation dataset is 200, with a total of 500 sentences evaluated. There are also 10 calibration sentences intended to detect potential misbehaviour.

DW02	0.50		
DW03	0.39	0.38	
DW04	0.33	0.36	0.45
	DW01	DW02	DW03

Table 48: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Macedonian and score Q1. The overall Krippendorff's alpha among all annotators is 0.40.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 48. The pairwise inter-annotator agreements for the score Q2 are shown in Table 49.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 50. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 51.

DW02	0.58		
DW03	0.33	0.49	
DW04	0.29	0.28	0.30
	DW01	DW02	DW03

Table 49: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Macedonian and score Q2. The overall Krippendorff's alpha among all annotators is 0.36.

DW01	DW02	DW03	DW04
1.00	0.96	1.00	0.55

Table 50: Calibration agreement computed via Krippendorff's alpha for Macedonian and score Q1.

DW01	DW02	DW03	DW04
1.00	-0.18	-0.03	-0.07

Table 51: Calibration agreement computed via Krippendorff's alpha for Macedonian and score Q2.

DW01	DW02	DW03	DW04
[74.96, 88.12]	[85.34, 91.92]	[78.26, 87.90]	[80.92, 89.16]

Table 52: 95% confidence intervals for the true mean of the score Q1 for Macedonian.

DW01	DW02	DW03	DW04
[82.90, 91.76]	[83.17, 89.71]	[76.93, 86.29]	[79.54, 88.02]

Table 53: 95% confidence intervals for the true mean of the score Q2 for Macedonian.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each evaluator are shown in Table 52. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 53.

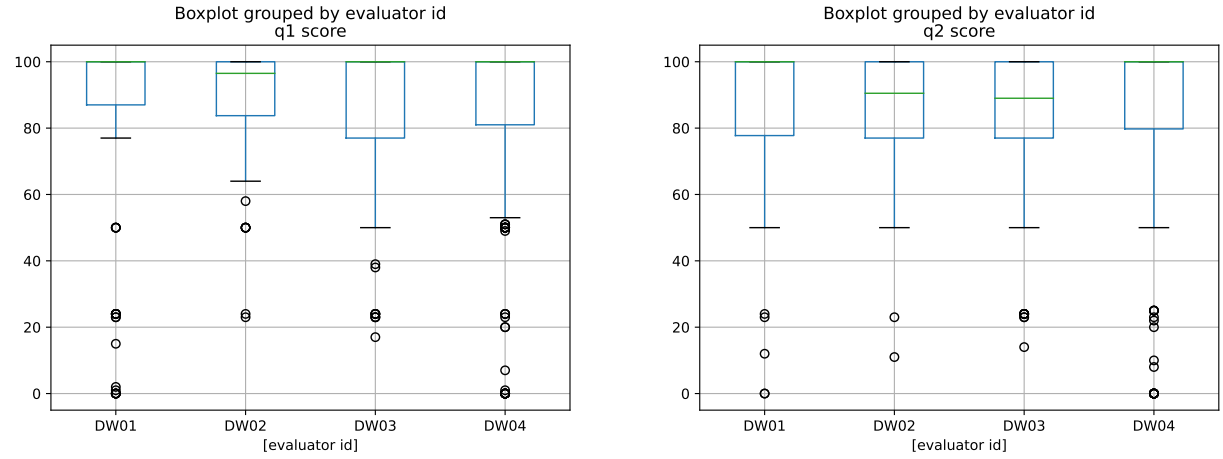


Figure 16: Boxplot of Q1 (left) and Q2 (right) scores for each annotator of English→Macedonian

Figure 16 shows a boxplot of the Q1 and Q2 scores per annotator.

The boxes are aligned, of comparable range, and highly rated.

The main issue that assessors identified with the machine translations was:

Punctuation One user commented: 'There were only 4-5 sentences that didn't make sense, but most recurring thing was false placement of punctuation marks.'

A second user commented: 'Grammatical errors with punctuation marks, such as quotation marks, are often repeated.' A third user commented that 'punctuation marks are not always in place'.

5.1.10 Hausa

Status: Silver standard, completed

This section contains the results of the human direct assessment of the English→Hausa system.

The total number of regular sentences in the evaluation dataset is 200, eight annotators have scored 1200 sentences. There are also 10 calibration sentences intended to detect potential misbehaviour.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 54. The pairwise inter-annotator agreements for the score Q2 are shown in Table 55.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 56. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 57.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each evaluator are shown in Table 58. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 59.

BBC02	0.67						
BBC03	0.66	0.95					
BBC04	0.61	0.75	0.72				
DW01	-0.42	-0.25	-0.24	-0.21			
DW02	-0.48	-0.37	-0.35	-0.27	-0.03		
DW03	-0.45	-0.29	-0.27	-0.21	0.20	0.41	
DW04	-0.31	-0.17	-0.15	-0.09	0.00	0.26	0.12
	BBC01	BBC02	BBC03	BBC04	DW01	DW02	DW03

Table 54: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Hausa and score Q1. The overall Krippendorff's alpha among all annotators is 0.18.

BBC02	0.57						
BBC03	0.61	0.95					
BBC04	0.45	0.49	0.45				
DW01	0.28	0.31	0.32	0.13			
DW02	-0.17	-0.04	-0.01	-0.36	0.18		
DW03	-0.11	0.02	0.06	-0.30	0.40	0.43	
DW04	0.20	0.20	0.26	0.04	0.51	0.48	0.23
	BBC01	BBC02	BBC03	BBC04	DW01	DW02	DW03

Table 55: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Hausa and score Q2. The overall Krippendorff's alpha among all annotators is 0.36.

BBC01	BBC02	BBC03	BBC04	DW01	DW02	DW03	DW04
-0.06	-0.11	-0.07	-0.13	-0.15	-0.19	-0.10	-0.09

Table 56: Calibration agreement computed via Krippendorff's alpha for Hausa and score Q1.

BBC01	BBC02	BBC03	BBC04	DW01	DW02	DW03	DW04
-0.16	0.00	0.00	-0.01	-0.19	-0.16	-0.17	-0.34

Table 57: Calibration agreement computed via Krippendorff's alpha for Hausa and score Q2.

BBC01	BBC02	BBC03	BBC04
[53.52, 65.18]	[53.03, 69.47]	[55.11, 71.41]	[63.99, 77.93]
DW01	DW02	DW03	DW04
[86.67, 91.98]	[88.97, 93.42]	[81.74, 88.71]	[90.25, 92.83]

Table 58: 95% confidence intervals for the true mean of the score Q1 for Hausa.

BBC01	BBC02	BBC03	BBC04
[69.58, 79.36]	[66.71, 82.13]	[64.58, 80.26]	[80.92, 89.90]
DW01	DW02	DW03	DW04
[39.63, 50.15]	[36.29, 46.49]	[18.97, 27.86]	[49.81, 59.12]

Table 59: 95% confidence intervals for the true mean of the score Q2 for Hausa.

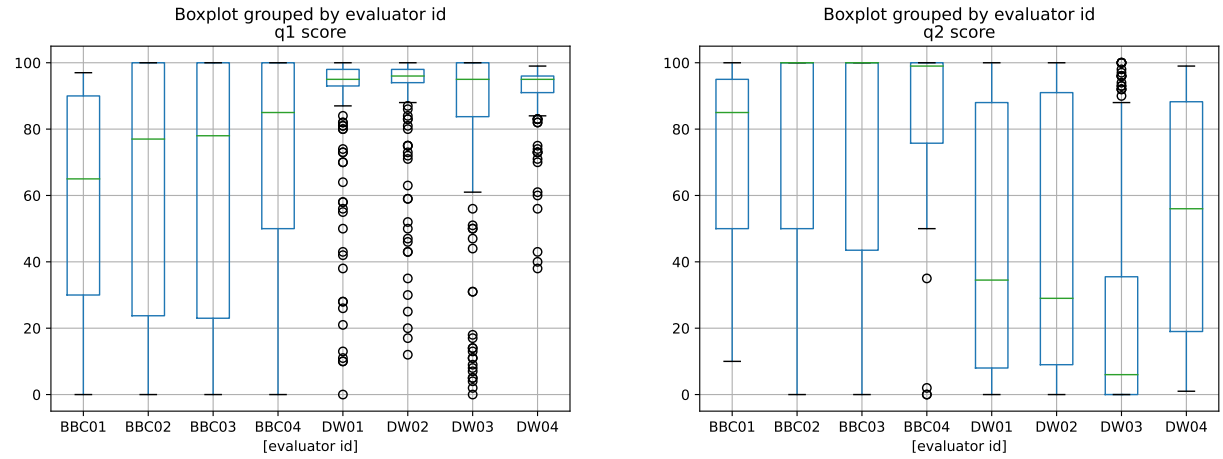


Figure 17: Boxplot of Q1 (left) and Q2 (right) scores for each annotator of English→Hausa

Figure 17 shows a boxplot of the Q1 and Q2 scores per annotator.

One interesting pattern is that while the evaluators from within one organisation scored sentences in a similar pattern with their colleagues (i.e. within just the BBC or DW), their appears to be a very different perception between organisations (i.e. comparing the BBC with DW). DW annotators' plots for fluency are also considerably lower than their adequacy scores.

The main issues that assessors identified with the machine translations were:

Gender One user commented that 'there are places where the ... sentences used different gender for the same subject' and another that 'there are gender differences in so many cases/sentences which I evaluated'.

Grammar One user commented: 'The translation has been amazing especially the one done by machines. I think it is generally okay minus some grammatical errors and gender mixture' but by contrast another commented: 'In some instances, the machine translation was *more* accurate in terms of grammar and semantics.' [our emphasis]

Meaning One user commented: 'In some place mistranslation of one or two words that would distort the meaning completely.'

5.1.11 Igbo

Status: Silver standard, completed

This section contains the results of the human direct assessment of the English→Igbo system.

The total number of regular sentences in the evaluation dataset is 200. There are also 10 calibration sentences intended to detect potential misbehaviour.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 60. The pairwise inter-annotator agreements for the score Q2 are shown in Table 61.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 62. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 63.

BBC02	0.61			
BBC03	0.41	0.49		
BBC04	0.69	0.54	0.49	
DW01	1.00	0.52	0.42	0.69
	BBC01	BBC02	BBC03	BBC04

Table 60: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Igbo and score Q1. The overall Krippendorff's alpha among all annotators is 0.52.

BBC02	0.53			
BBC03	0.55	0.28		
BBC04	0.58	0.36	0.31	
DW01	1.00	0.40	0.35	0.61
	BBC01	BBC02	BBC03	BBC04

Table 61: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Igbo and score Q2. The overall Krippendorff's alpha among all annotators is 0.42.

BBC01	BBC02	BBC03	BBC04	DW01
0.97	1.00	0.90	1.00	0.89

Table 62: Calibration agreement computed via Krippendorff's alpha for Igbo and score Q1.

BBC01	BBC02	BBC03	BBC04	DW01
-0.05	-0.06	-0.15	-0.10	-0.05

Table 63: Calibration agreement computed via Krippendorff's alpha for Igbo and score Q2.

BBC01	BBC02	BBC03	BBC04	DW01
[62.92, 72.76]	[58.85, 67.05]	[54.33, 65.79]	[63.38, 71.28]	[59.63, 69.97]

Table 64: 95% confidence intervals for the true mean of the score Q1 for Igbo.

BBC01	BBC02	BBC03	BBC04	DW01
[70.78, 81.22]	[69.70, 77.88]	[71.47, 79.04]	[60.20, 68.79]	[62.64, 74.12]

Table 65: 95% confidence intervals for the true mean of the score Q2 for Igbo.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each evaluator are shown in Table 64. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 65.

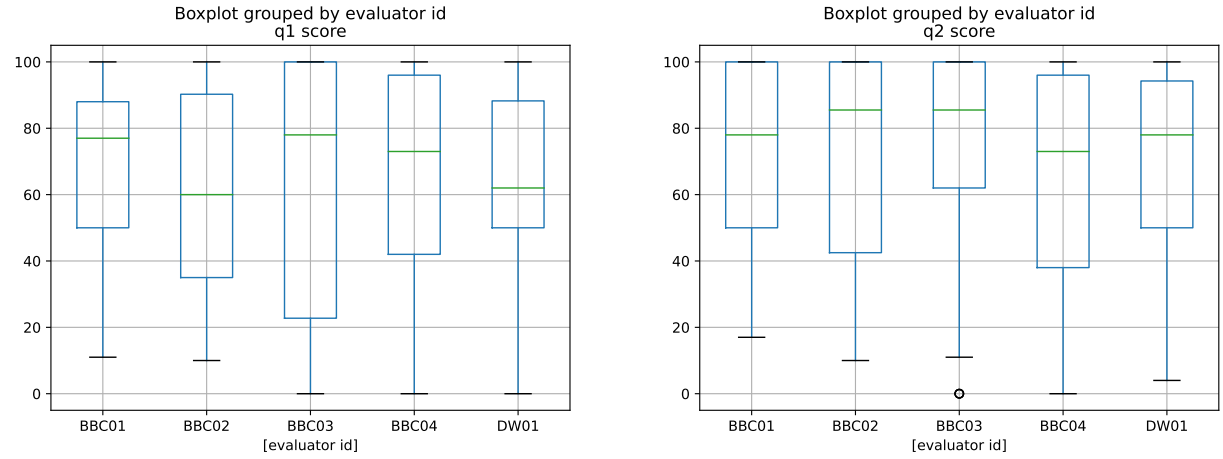


Figure 18: Boxplot of Q1 (left) and Q2 (right) scores for each annotator of English→Igbo

Figure 18 shows a boxplot of the Q1 and Q2 scores per annotator.

It is worth noting that while scores for Igbo were on the lower side of the scale for BLEU, spBLEU and COMET scores, the FLORES scores suggested higher ranges of usability. This appears to have been reflected in the adequacy and fluency scores of the evaluators all of which have high median scores.

The main issues that assessors identified with the machine translations were:

Grammar One user commented: 'The machine tries but in some cases, it is obvious that it is not the proper manner of speaking.' They conclude: 'But it is a good effort. Well done, Machine.'

Meaning One user commented: 'Some of the sentences are not well structured thereby having different meaning from the actual meaning.' Also: 'There were instances where the names of places were changed which is passes false information and doesn't represent the true motive of the translation.'

Punctuation One user commented that 'there are punctuation errors in some of the sentences, although not in all. It will require extra carefulness to read, understand and comprehend them, and might be difficult for someone without deep knowledge of the language.'

5.1.12 Tigrinya

Status: Silver standard, completed

This section contains the results of the human direct assessment of the English→Tigrinya system.

The total number of regular sentences in the evaluation dataset is 200. There are also 10 calibration sentences intended to detect potential misbehaviour.

BBC02	0.24				
BBC03	0.43	1.00			
BBC04	0.35	0.36	0.33		
DW02	0.41	1.00	0.10	0.43	
DW03	0.45	0.26	1.00	0.21	1.00
	BBC01	BBC02	BBC03	BBC04	DW02

Table 66: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Tigrinya and score Q1. The overall Krippendorff's alpha among all annotators is 0.36.

BBC02	0.27				
BBC03	0.52	1.00			
BBC04	0.42	0.40	0.71		
DW02	0.32	1.00	0.15	0.32	
DW03	0.53	0.29	1.00	0.37	1.00
	BBC01	BBC02	BBC03	BBC04	DW02

Table 67: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Tigrinya and score Q2. The overall Krippendorff's alpha among all annotators is 0.42.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 66. The pairwise inter-annotator agreements for the score Q2 are shown in Table 67.

BBC01	BBC02	BBC03	BBC04	DW02	DW03
-0.16	-0.06	-0.12	-0.23	0.07	-0.12

Table 68: Calibration agreement computed via Krippendorff's alpha for Tigrinya and score Q1.

BBC01	BBC02	BBC03	BBC04	DW02	DW03
-0.11	-0.25	1.00	-0.17	-0.39	1.00

Table 69: Calibration agreement computed via Krippendorff's alpha for Tigrinya and score Q2.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 68. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 69.

BBC01	BBC02	BBC03	BBC04	DW02	DW03
[37.03, 45.09]	[40.71, 54.31]	[25.17, 42.37]	[58.72, 68.68]	[60.09, 68.87]	[17.48, 32.84]

Table 70: 95% confidence intervals for the true mean of the score Q1 for Tigrinya.

BBC01	BBC02	BBC03	BBC04	DW02	DW03
[42.57, 50.51]	[45.82, 59.18]	[29.38, 47.42]	[45.19, 56.07]	[62.29, 70.09]	[30.61, 47.09]

Table 71: 95% confidence intervals for the true mean of the score Q2 for Tigrinya.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each

evaluator are shown in Table 70. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 71.

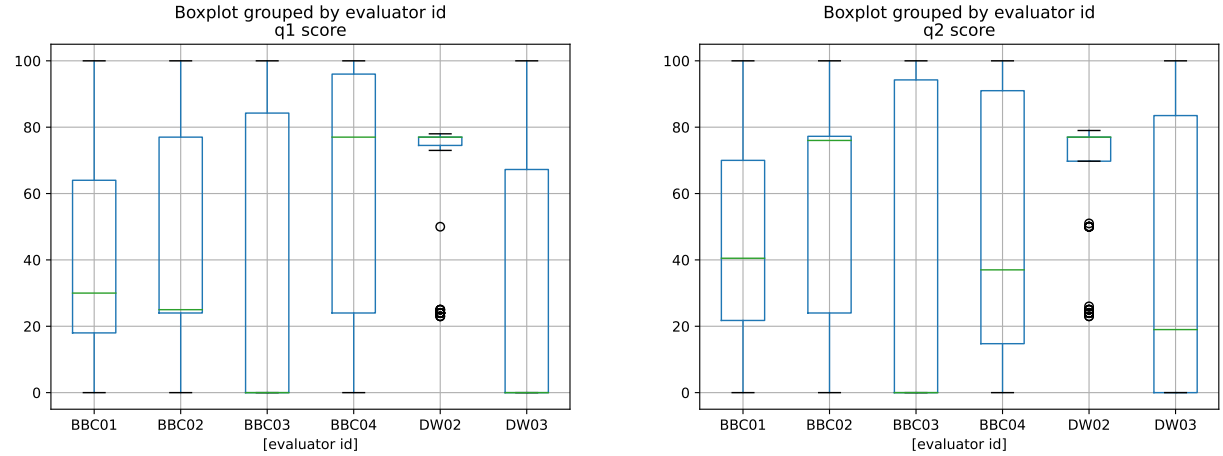


Figure 19: Boxplot of Q1 (left) and Q2 (right) scores for each annotator of English→Tigrinya

Figure 19 shows a boxplot of the Q1 and Q2 scores per annotator.

The charts for Tigrinya are notable for their extended range except for the evaluator DW02, indicating the perceived quality of sentences were not consistent across the board.

The main issues that assessors identified with the machine translations included:

Meaning One user commented: 'Some sentences are good while others are totally distorted in the translation and require a lot of work.'

One cause of confused meaning that was repeatedly mentioned was ommission:

Omission One user commented: 'There are incomplete sentences and the quality of translation was not up to the standard.' Another user commented: 'I have noticed there are incomplete sentences and they need to be completed.'

5.1.13 Pashto

Status: Silver standard, completed

This section contains the results of the human direct assessment of the English→Pashto system.

The total number of regular sentences in the evaluation dataset is 200. There are also 10 calibration sentences intended to detect potential misbehaviour.

As the language selected for the 'surprise language' challenge, Pashto was extensively evaluated by both BBC and DW.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 72. The pairwise inter-annotator agreements for the score Q2 are shown in Table 73.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 74. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 75.

BBC02	0.29			
BBC04	0.34	0.10		
DW01	0.36	0.14	0.53	
DW02	0.27	0.21	0.12	0.10
	BBC01	BBC02	BBC04	DW01

Table 72: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Pashto and score Q1. The overall Krippendorff's alpha among all annotators is 0.30.

BBC02	0.38			
BBC04	0.26	0.19		
DW01	0.34	0.29	0.56	
DW02	0.40	0.26	-0.07	0.01
	BBC01	BBC02	BBC04	DW01

Table 73: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Pashto and score Q2. The overall Krippendorff's alpha among all annotators is 0.31.

BBC01	BBC02	BBC04	DW01	DW02
0.99	0.99	0.77	0.85	0.86

Table 74: Calibration agreement computed via Krippendorff's alpha for Pashto and score Q1.

BBC01	BBC02	BBC04	DW01	DW02
-0.04	-0.01	-0.43	-0.10	1.00

Table 75: Calibration agreement computed via Krippendorff's alpha for Pashto and score Q2.

BBC01	BBC02	BBC04	DW01	DW02
[76.83, 85.12]	[88.42, 93.64]	[66.00, 72.53]	[61.09, 68.84]	[81.38, 85.66]

Table 76: 95% confidence intervals for the true mean of the score Q1 for Pashto.

BBC01	BBC02	BBC04	DW01	DW02
[80.59, 88.08]	[81.45, 87.49]	[58.84, 65.83]	[59.19, 68.01]	[89.34, 93.70]

Table 77: 95% confidence intervals for the true mean of the score Q2 for Pashto.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each evaluator are shown in Table 76. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 77.

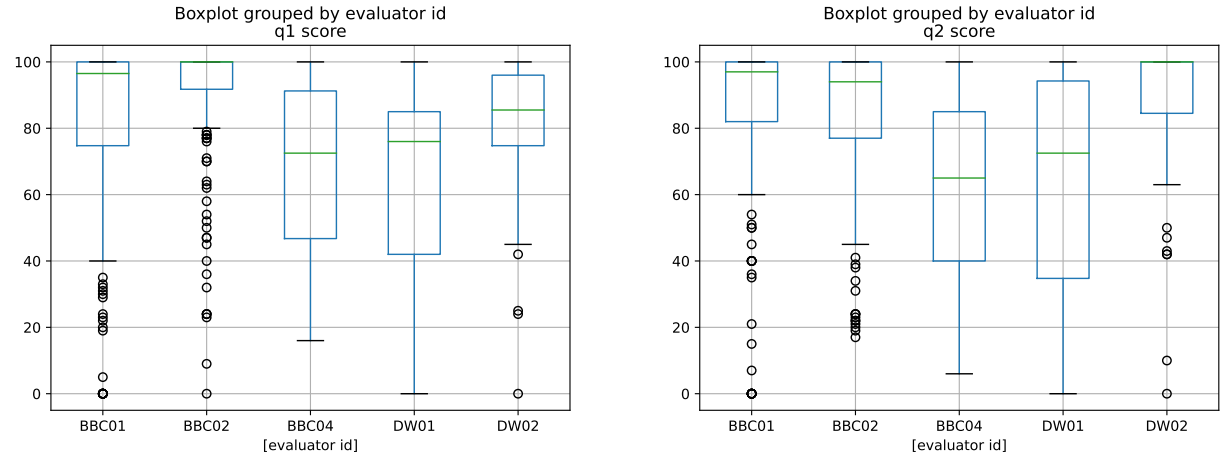


Figure 20: Boxplot of Q1 (left) and Q2 (right) scores for each annotator of English→Pashto

Figure 20 shows a boxplot of the Q1 and Q2 scores per annotator.

The main issues that assessors identified with the machine translations were:

Alphabet One user commented: 'There is a problem with Latin and Arabic (Pashto) fonts in the same sentences.' Another user commented: 'The sentences were not shown correctly, if other symbols/letters than Pashto were included in Pashto sentence.'

Meaning Common issues included 'occassional addition or deletion of details like dates and days' and 'confusion of adverbs and determiners (e.g. "more than 7 thousand killed" vs. "almost 7 thousand killed")'.

5.1.14 Burmese

Status: Bronze standard, completed

This section contains the results of the human direct assessment of the English→Burmese system.

The total number of regular sentences in the evaluation dataset is 200. There are also 10 calibration sentences intended to detect potential misbehaviour.

BBC02	0.15		
BBC03	0.13	0.70	
DW02	0.12	1.00	0.41
	BBC01	BBC02	BBC03

Table 78: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Burmese and score Q1. The overall Krippendorff's alpha among all annotators is 0.42.

BBC02	0.10		
BBC03	0.14	0.62	
DW02	0.11	1.00	0.48
	BBC01	BBC02	BBC03

Table 79: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Burmese and score Q2. The overall Krippendorff's alpha among all annotators is 0.41.

BBC01	BBC02	BBC03	DW02
0.41	0.99	0.79	0.91

Table 80: Calibration agreement computed via Krippendorff's alpha for Burmese and score Q1.

BBC01	BBC02	BBC03	DW02
-0.76	-0.20	-0.18	-0.48

Table 81: Calibration agreement computed via Krippendorff's alpha for Burmese and score Q2.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 78. The pairwise inter-annotator agreements for the score Q2 are shown in Table 79.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 80. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 81.

BBC01	BBC02	BBC03	DW02
[39.87, 42.47]	[24.83, 41.01]	[19.08, 28.44]	[21.67, 36.01]

Table 82: 95% confidence intervals for the true mean of the score Q1 for Burmese.

BBC01	BBC02	BBC03	DW02
[42.86, 45.53]	[23.13, 39.11]	[27.24, 37.43]	[18.85, 32.69]

Table 83: 95% confidence intervals for the true mean of the score Q2 for Burmese.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each evaluator are shown in Table 82. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 83.

Figure 21 shows a boxplot of the Q1 and Q2 scores per annotator.

The boxplots for both adequacy and fluency are on the higher end of the scale for Burmese. This is noteworthy as the Burmese scores in the automated evaluations (BLEU, spBLEU, FLORES, COMET) did not suggest such positive outcomes except for the chrF scores where GoURMET appeared to fare marginally better than Google in this direction.

The main issues that assessors identified with the machine translations were:

Grammar One user commented: 'Sometimes the whole sentence meaning is wrong because of only one preposition, which is because Burmese and English sentence structure are different.'

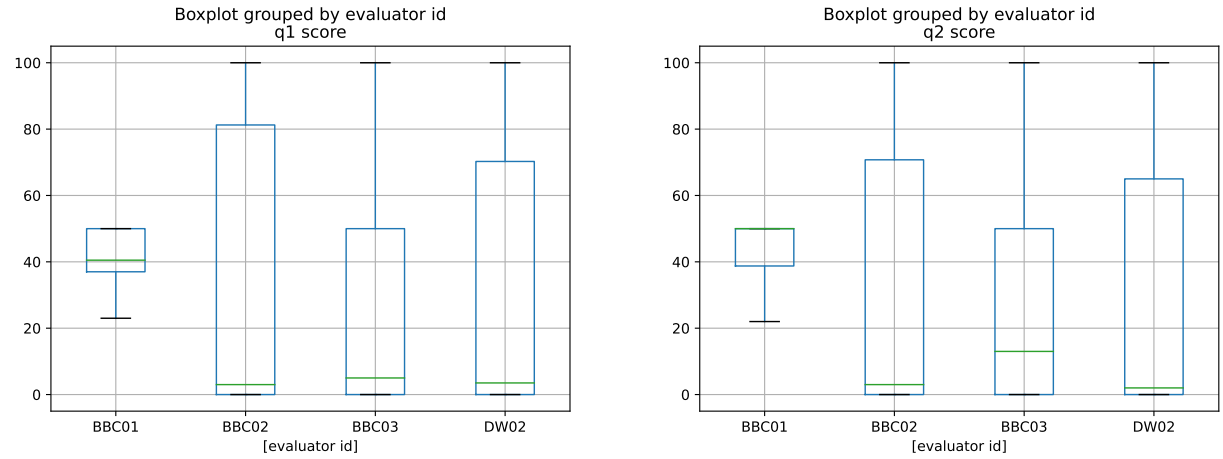


Figure 21: Boxplot of Q1 (left) and Q2 (right) scores for each annotator of English→Burmese

Sentence length One user commented: 'It appears sentences almost match and are likely to be more authentic when sentences are very short.'

5.1.15 Yoruba

Status: Silver standard, completed

This section contains the results of the human direct assessment of the English→Yoruba system.

The total number of regular sentences in the evaluation dataset is 200. There are also 10 calibration sentences intended to detect potential misbehaviour.

BBC02	0.44		
BBC04	0.17	0.21	
DW03	-0.40	-0.08	-0.48
	BBC01	BBC02	BBC04

Table 84: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Yoruba and score Q1. The overall Krippendorff's alpha among all annotators is 0.11.

BBC02	0.17		
BBC04	0.07	-0.18	
DW03	0.06	0.42	-0.16
	BBC01	BBC02	BBC04

Table 85: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Yoruba and score Q2. The overall Krippendorff's alpha among all annotators is 0.14.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 84. The pairwise inter-annotator agreements for the score Q2 are shown in Table 85.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 86. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 87.

BBC01	BBC02	BBC04	DW03
1.00	1.00	0.61	0.97

Table 86: Calibration agreement computed via Krippendorff's alpha for Yoruba and score Q1.

BBC01	BBC02	BBC04	DW03
-0.02	1.00	-0.21	0.00

Table 87: Calibration agreement computed via Krippendorff's alpha for Yoruba and score Q2.

BBC01	BBC02	BBC04	DW03
[7.35, 14.92]	[19.95, 35.95]	[19.01, 23.49]	[61.14, 69.47]

Table 88: 95% confidence intervals for the true mean of the score Q1 for Yoruba.

BBC01	BBC02	BBC04	DW03
[25.70, 37.71]	[58.61, 74.79]	[22.68, 28.90]	[54.05, 64.78]

Table 89: 95% confidence intervals for the true mean of the score Q2 for Yoruba.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each evaluator are shown in Table 88. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 89.

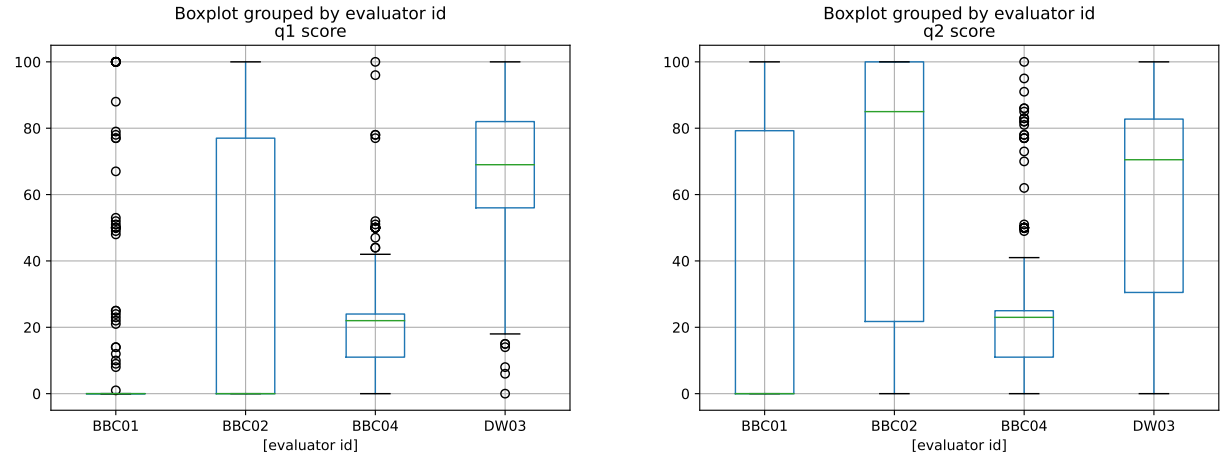


Figure 22: Boxplot of Q1 (left) and Q2 (right) scores for each annotator of English→Yoruba

Figure 22 shows a boxplot of the Q1 and Q2 scores per annotator.

The scores include some of the lowest ratings across all the languages in scope. It is notable however, that the DW evaluator's scores are distinctly different from the BBC evaluators' scores.

The main issue that assessors identified with the machine translations:

Spelling One user commented: 'The machine translation was not bad in its entirety. It did a fantastic job, but for little spelling mistakes/errors. But the errors do not really affect the general meaning of the sentences.'

5.1.16 Urdu

Status: Silver standard, completed

This section contains the results of the human direct assessment of the English→Urdu system.

The total number of regular sentences in the evaluation dataset is 200. There are also 10 calibration sentences intended to detect potential misbehaviour.

BBC02	0.69					
BBC03	0.45	1.00				
BBC04	0.31	0.32	0.42			
DW01	0.31	1.00	0.41	0.27		
DW03	0.46	0.44	1.00	0.40	1.00	
OTHER	0.81	0.77	0.44	0.27	0.28	0.42
	BBC01	BBC02	BBC03	BBC04	DW01	DW03

Table 90: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Urdu and score Q1. The overall Krippendorff's alpha among all annotators is 0.51.

BBC02	0.71					
BBC03	0.50	1.00				
BBC04	0.20	0.19	0.25			
DW01	0.74	1.00	0.25	0.13		
DW03	0.53	0.44	1.00	-0.01	1.00	
OTHER	0.43	0.34	0.10	-0.29	0.51	-0.01
	BBC01	BBC02	BBC03	BBC04	DW01	DW03

Table 91: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Urdu and score Q2. The overall Krippendorff's alpha among all annotators is 0.31.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 90. The pairwise inter-annotator agreements for the score Q2 are shown in Table 91.

BBC01	BBC02	BBC03	BBC04	DW01	DW03	OTHER
0.53	0.94	0.26	-0.08	1.00	0.41	0.78

Table 92: Calibration agreement computed via Krippendorff's alpha for Urdu and score Q1.

BBC01	BBC02	BBC03	BBC04	DW01	DW03	OTHER
-0.57	-0.21	-0.74	-0.76	1.00	-0.62	-0.32

Table 93: Calibration agreement computed via Krippendorff's alpha for Urdu and score Q2.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 92. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 93.

BBC01	BBC02	BBC03	BBC04	DW01	DW03	OTHER
[35.51, 45.31]	[27.40, 42.18]	[29.90, 36.02]	[13.02, 19.18]	[20.17, 43.75]	[27.75, 33.75]	[37.44, 47.18]

Table 94: 95% confidence intervals for the true mean of the score Q1 for Urdu.

BBC01	BBC02	BBC03	BBC04	DW01	DW03	OTHER
[30.83, 40.29]	[22.69, 37.19]	[29.39, 35.45]	[6.89, 11.85]	[29.33, 59.15]	[27.77, 33.79]	[56.06, 64.96]

Table 95: 95% confidence intervals for the true mean of the score Q2 for Urdu.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each evaluator are shown in Table 94. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 95.

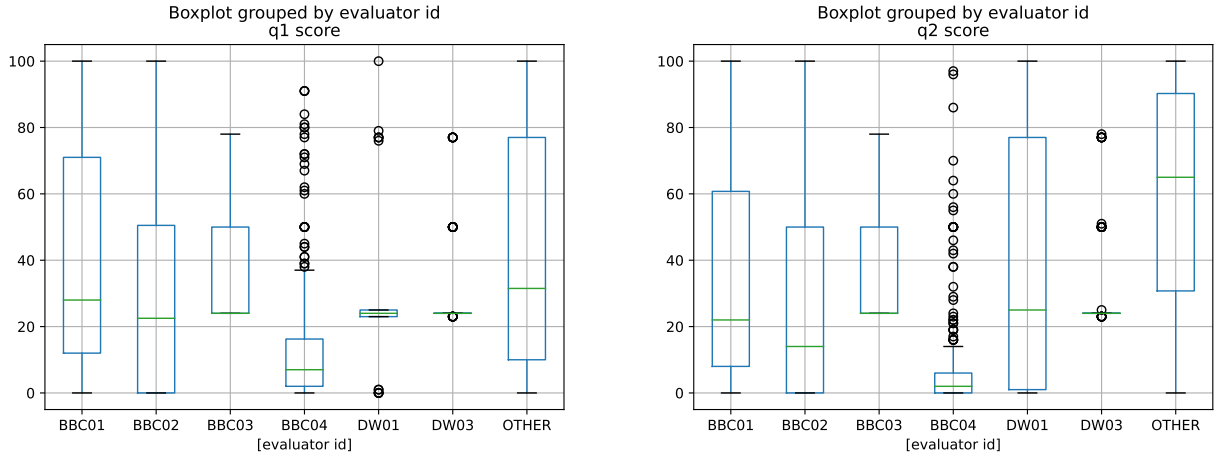


Figure 23: Boxplot of Q1 (left) and Q2 (right) scores for each annotator of English→Urdu

Figure 23 shows a boxplot of the Q1 and Q2 scores per annotator.

The scores for Urdu are less impressive than other languages, with the median scores towards the bottom of most boxplots. This is also echoed in the comments gathered in the post-evaluation exercise. However, Urdu appears among the top models in the automated evaluation in terms of its spBLEU, COMET and chrF scores. With further time it would be interesting to investigate potential reasons for this divergence.

The main issues that assessors identified with the machine translations included:

Gender One user commented: 'There were many issues with the translations; gender was the most prominent among them.' Another user commented: 'Difference of gender is not fully recognised.'

Grammar One user commented: 'To me it seems that the most of the translations were too mechanical and literal. Most of them were wrong in grammar.'

Sentence length One user commented that 'most of the sentences were grammatically wrong and the longer the sentence, there were more chances of mistakes'. Another user commented: 'Long sentences are mostly wrong but short sentences make more sense.'

Tense One user commented that 'in a couple of instances the tense wasn't clear enough, even for short sentences' and another that 'I think active and passive sentences were also a problem'.

5.1.17 Turkish v2

Status: Silver standard, completed

The total number of regular sentences in the evaluation dataset is 200. There are also 10 calibration sentences intended to detect potential misbehaviour.

BBC03	0.60			
BBC04	0.53	0.68		
DW01	0.37	0.41	0.41	
DW02	0.41	0.48	0.47	0.38
	BBC02	BBC03	BBC04	DW01

Table 96: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Turkish v2 and score Q1. The overall Krippendorff's alpha among all annotators is 0.49.

BBC03	0.56			
BBC04	0.62	0.66		
DW01	0.48	0.48	0.47	
DW02	0.46	0.55	0.55	0.28
	BBC02	BBC03	BBC04	DW01

Table 97: Pairwise inter-annotator agreement computed via Krippendorff's alpha for Turkish v2 and score Q2. The overall Krippendorff's alpha among all annotators is 0.52.

The pairwise inter-annotator agreements for the score Q1 measured via Krippendorff's alpha for *interval* data are shown in Table 96. The pairwise inter-annotator agreements for the score Q2 are shown in Table 97.

BBC02	BBC03	BBC04	DW01	DW02
1.00	1.00	1.00	1.00	0.93

Table 98: Calibration agreement computed via Krippendorff's alpha for Turkish v2 and score Q1.

BBC02	BBC03	BBC04	DW01	DW02
-0.20	-0.20	-0.11	-0.22	-0.08

Table 99: Calibration agreement computed via Krippendorff's alpha for Turkish v2 and score Q2.

The calibration agreements for the adequacy score (Q1) measured via Krippendorff's alpha are shown in Table 98. This measure can be considered as a proxy for misbehaviour detection. The calibration agreements for the fluency score (Q2) are shown in Table 99.

The mean of each annotator's scores is an overall indicator of the NMT system performance. The 95% confidence intervals assuming a normal distribution for the mean of the score Q1 for each evaluator are shown in Table 100. The 95% confidence intervals for the mean of the score Q2 for each evaluator are shown in Table 101.

BBC02	BBC03	BBC04	DW01	DW02
[54.35, 63.59]	[59.76, 69.12]	[66.90, 76.40]	[68.27, 76.66]	[69.55, 75.35]

Table 100: 95% confidence intervals for the true mean of the score Q1 for Turkish v2.

BBC02	BBC03	BBC04	DW01	DW02
[66.32, 74.77]	[66.44, 76.03]	[66.80, 75.93]	[51.98, 61.29]	[72.89, 78.90]

Table 101: 95% confidence intervals for the true mean of the score Q2 for Turkish v2.

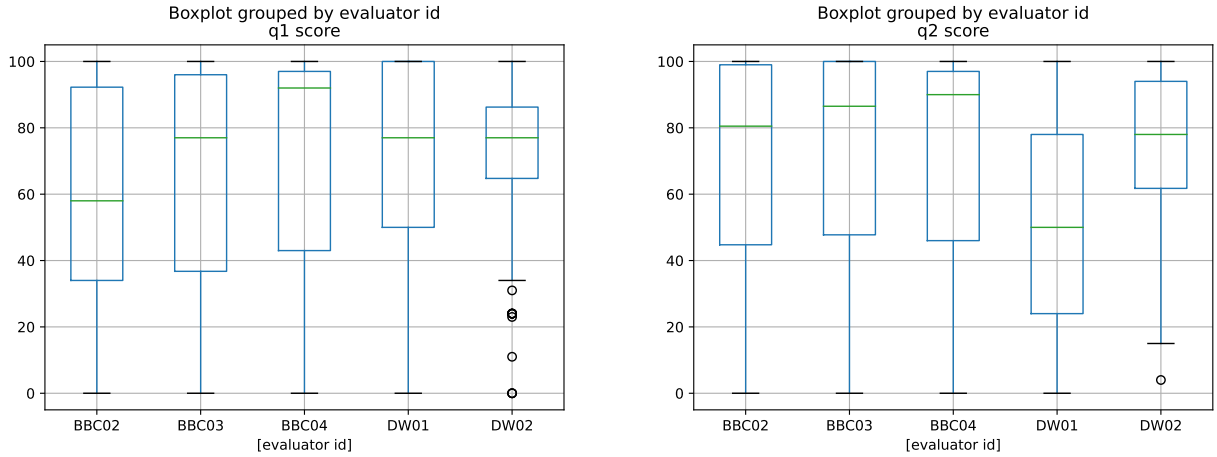


Figure 24: Boxplot of Q1 (left) and Q2 (right) scores for each annotator of English→Turkish v2

Figure 24 shows a boxplot of the Q1 and Q2 scores per annotator.

Following the work to improve the original Turkish translation model as Turkish v2 in the second half of the project, it is now one of the highest scoring languages for direct assessment. The scores of most annotators are also broadly consistent, with scores for both adequacy (Q1) and fluency (Q2) generating similar boxplots.

The main issues that assessors identified with the machine translations were:

Grammar One user commented: 'Because of the Turkish grammar and syntax structure, the machine is still coming up with major grammatical errors in complex sentences.'

Phrasing One user commented: 'Some of the sentences are well translated by machine, sometimes nearly 100%. But some of them don't have a common meaning at all and it's usually because of the Turkish expressions.'

Proper nouns One user commmented: 'There are problems with special names like movie titles, which the system kept in English.

5.1.18 Direct Assessment Evaluation Findings

Direct assessment evaluations were conducted by over 100 journalists at the BBC and DW. Two series of evaluation sets were compiled, each with 110 sentences, from test sets in every GoURMET language. Provided they broadcast in that language, each media partner recruited and briefed up to five evaluators, with each team given a window of around six to eight weeks to complete the assessments.

Three languages (**Gujarati**, **Macedonian** and **Burmese**) had fewer annotations and so achieved 'bronze' standard. The lower number of annotations can be linked to the fact that these languages were served by only one, not both, of the media partners. The remaining languages were conducted to 'silver' standard, with over 600 annotations each.

Of the thirteen languages at 'silver' standard, three languages (**Urdu**, **Serbian** and **Turkish**) went on to be evaluated to 'gold' standard, which involved a further round of post-editing assessment as described in section 2.3 with results in section 6.1.

A key consideration for human evaluation is working out what level of competency is '*good enough*'. This depends very much on the target application and context. The standards of accuracy required in a news setting are among the highest, particularly for content that will be seen by audiences (i.e. the 'content creation' use case), with arguably more room for flexibility in the 'media monitoring' use case.

It should be noted that in a few cases the inter-annotator scores were divergent between BBC and DW evaluators (e.g. Hausa in section 5.1.10). There seem to be different stylistic preferences at the respective organisations. We are also mindful that it is considered 'difficult for human evaluators to completely isolate fluency from adequacy' (Callison-Burch et al., 2007).

It has proved helpful to invite evaluators to provide direct comments at the end of the evaluation sets. These offer valuable direct feedback for media partners and their internal stakeholders in particular.

Regardless of the language, the most common point that came across in the direct feedback was the observation that systems deal much better with shorter sentences, rather than long, complicated ones. However, there were very few short, simple sentences in the evaluation sets with most sentences including more than 10-12 words. If could be argued that the assessment was a particularly challenging one for this reason.

The translation of gender, proper nouns and subject/verb agreements were also among issues frequently raised by evaluators as areas that need further enhancement.

Serbian has the most positive outcomes of all the direct assessment results. Among the languages that have scored well in automated evaluation (e.g. spBLEU and COMET metrics) we see that Pashto and Turkish have drawn even higher scores from human evaluators, Hausa, Amharic and Igbo have fared well, but Urdu and Burmese have not scored as highly.

There is one caveat that needs noting about the approach we have applied in direct assessments. The method assigns primacy to the professional translations, assuming them to be 'better' than MT output. The sentences in the test set are not custom translated by a professional translator but compiled from material translated by journalists and published on BBC and DW platforms. The evaluators have therefore been asked to compare machine translated sentences with a human translation (HT) that's treated as the 'ground truth' and are asked to state how much 'the black text [MT] adequately expresses the meaning of the grey text [HT]'.

Evaluators therefore consider the accuracy of the MT *not* against the source but a derivative of it in the target language translated by a fellow journalist for publishing. Having a closer look at our evaluation data, there are cases when this approach might have skewed the results. Below are some samples from the Turkish set with points to compare *in blue italic text*:

Original (reference) sentence For his part, Saudi Arabia's King *Salman has called for* "a decisive stance from the international community against Iran".

GoURMET output Suudi Arabistan Kralı *Salman* ise, "uluslararası toplumun İran'a karşı kararlı bir tutum sergilemesi" çağrısında *bulundu*.

Human translation Suudi Arabistan Kralı *Selman bin Abdülaziz el-Suud* ise "İran'a karşı uluslararası toplumdan kararlı bir duruş" çağrısı *yapıyor*.

Back translation Whereas Saudi Arabia's King *Salman bin Abdulaziz Al Saud is calling for* "a decisive stance from the international community against Iran".

For an evaluator just seeing the two translations, it would look like the MT is failing to convey the full name of the subject (King Salman bin Abdulaziz al-Suud) and that the tense is not rendered accurately (present vs past). However, compared with the original sentence in English, the MT output could have scored higher on adequacy as it renders the name and the tense faithfully, whereas the human translation takes liberties.

A similar, albeit more subtle example, is below, again with points to compare *in blue italic text*:

Original (reference) sentence Merkel was in charge for *18 years*.

GoURMET output Merkel *18 yıl* boyunca görevdeydi.

Human translation Merkel bu görevi *tam 18 yıl* boyunca sürdürdü.

Back translation Merkel has held this post for *a full 18 years*.

Again, the MT in this case is *not missing* information but the HT which contains subtle additions that MT cannot be expected to replicate.

With more time, it would be interesting to conduct deeper analysis of other evaluation sets with a translation expert to see whether this is a repetitive pattern across languages we have covered.

This is partly linked to the nature of the news domain. Bielsa and Bassnett note that among other devices, 'Contextualization… addition of new information, or elimination' are 'all part of the ordinary operations of news translation.' (Bassnett and Bielsa, 2008)

We also note that in *Taking MT Evaluation Metrics to Extremes*, Fomicheva and Specia (2019) suggest 'that the focus of future research on MT evaluation should move from handling acceptable variation between MT output and reference translations to estimating the impact of translation errors on MT quality.' Bearing in mind the particularities of reversioning and localization in news, it would be interesting for future projects to explore how this might be applied for the news domain and its use cases.

5.2 Gap Filling

This section describes the gap-filling results by language.

According to terms set out in deliverable D5.4 section 2.2.8 , the minimum number of evaluators required to complete the gap filling task is 6 to achieve the 'bronze' standard and 9 for the 'silver' standard.

BBC and DW have conducted this task on a best-effort basis, in line with the availability of the language in their respective portfolios. Each partner invited up to 9 evalutors per language. All of the languages included in the scope of the project between m1-42 were evaluated to at least a silver standard.

The gap-filling results are also combined without commentary, so they can be viewed at a glance in Appendix A:

The statistics for the GF evaluation for all languages into English are shown in Table 132.

The boxplots from Figures 25-41 are reproduced on page 79.

5.2.1 Swahili

Status: Silver standard, completed

The statistics for the GF evaluation for `sw→en` are shown in Table 102. The detailed boxplot of results is shown in Figure 25.

The boxes for Google and GoURMET clearly overlap, implying that the difference in usefulness is not significant.

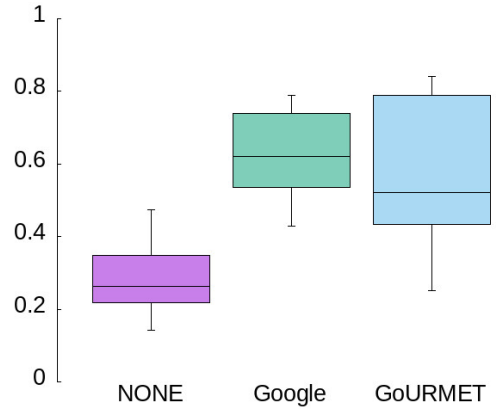


Figure 25: Results of GF evaluation `sw→en`

Unique Evaluators	18
Number of unique gaps	70
Average number of gaps per sentence	2.3
Evaluations per gap-configuration	6.3
Number of evaluations by hint type: NONE	448
Number of evaluations by hint type: GoURMET	439
Number of evaluations by hint type: Google	443

Table 102: Summary of GF evaluation `sw→en`

5.2.2 Gujarati

Status: Silver standard, completed

The statistics for the GF evaluation for gu→en are shown in Table 103. The detailed boxplot of results is shown in Figure 26.

There is an overlap between Google and GoURMET, although Google performs slightly better for this language pair overall.

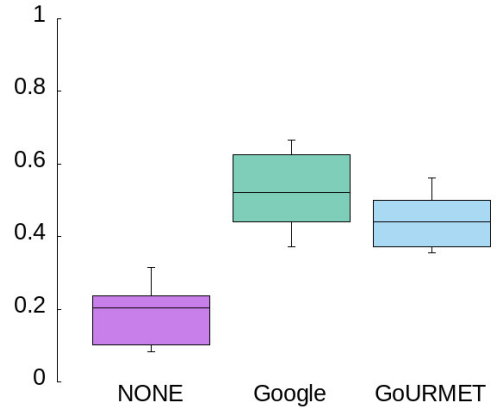


Figure 26: Results of GF evaluation gu→en

Unique Evaluators	15
Number of unique gaps	143
Average number of gaps per sentence	4.8
Evaluations per gap-configuration	5.1
Number of evaluations by hint type: NONE	725
Number of evaluations by hint type: GoURMET	725
Number of evaluations by hint type: Google	725

Table 103: Summary of GF evaluation gu→en

5.2.3 Turkish

Status: Silver standard, completed

The statistics for the GF evaluation for the first version of tr→en translation model are shown in Table 104. The detailed boxplot of results is shown in Figure 27.

The perceived quality appeared lower than for Google. This language pair was revisited later in the project as Turkish v2 (see section 5.2.17).

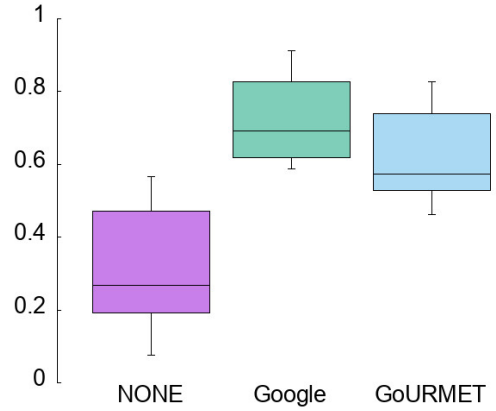


Figure 27: Results of GF evaluation tr→en v1

Unique Evaluators	18
Number of unique gaps	83
Average number of gaps per sentence	2.8
Evaluations per gap-configuration	6.0
Number of evaluations by hint type: NONE	498
Number of evaluations by hint type: GoURMET	498
Number of evaluations by hint type: Google	498

Table 104: Summary of GF evaluation tr→en v1

5.2.4 Bulgarian

Status: Silver standard, completed

The statistics for the GF evaluation for bg→en are shown in Table 105. The detailed boxplot of results is shown in Figure 28.

The boxplot for this language pair indicates broad similarity, and in this case the difference between Google and GoURMET is not significant.

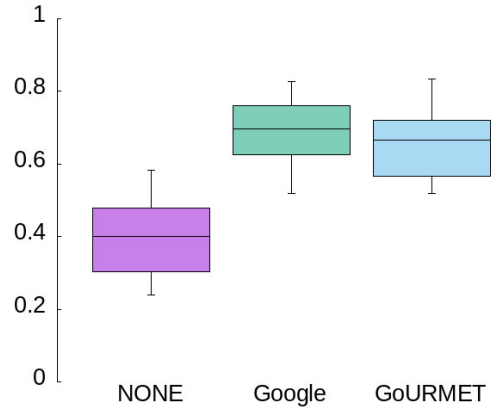


Figure 28: Results of GF evaluation bg→en

Unique Evaluators	19
Number of unique gaps	72
Average number of gaps per sentence	2.4
Evaluations per gap-configuration	6.3
Number of evaluations by hint type: NONE	457
Number of evaluations by hint type: GoURMET	456
Number of evaluations by hint type: Google	455

Table 105: Summary of GF evaluation bg→en

5.2.5 Tamil

Status: Silver standard, completed

The statistics for the GF evaluation for ta→en are shown in Table 106. The detailed boxplot of results is shown in Figure 29.

The boxplot for this language pair indicates broad similarity, with Google performing slightly better than GoURMET.

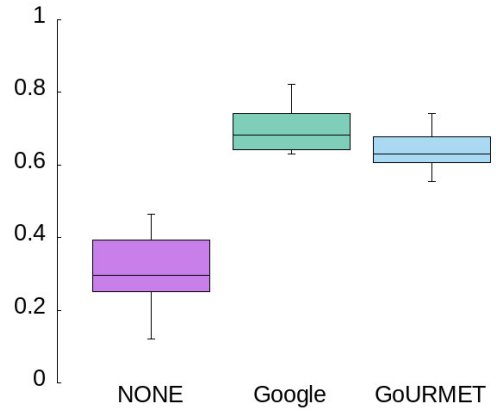


Figure 29: Results of GF evaluation ta→en

Unique Evaluators	13
Number of unique gaps	95
Average number of gaps per sentence	3.2
Evaluations per gap-configuration	4.4
Number of evaluations by hint type: NONE	425
Number of evaluations by hint type: GoURMET	412
Number of evaluations by hint type: Google	411

Table 106: Summary of GF evaluation ta→en

5.2.6 Serbian

Status: Silver standard, completed

The statistics for the GF evaluation for sr→en are shown in Table 107. The detailed boxplot of results is shown in Figure 30.

Results across evaluators are significantly close and high for Serbian, as illustrated in the ranges of the boxes. The boxplot for this language pair is among the highest of all GoURMET languages, along with Macedonian, suggesting a robust degree of usability. However, the figures still remain marginally below Google's figures.

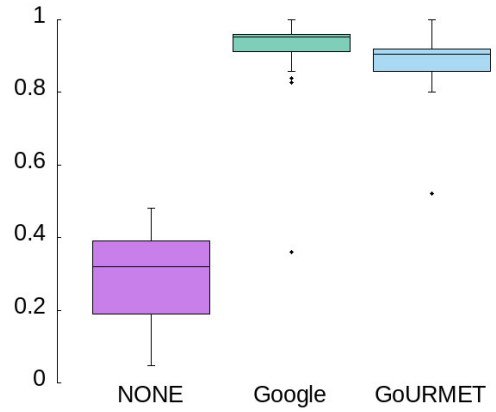


Figure 30: Results of GF evaluation sr→en

Unique Evaluators	17
Number of unique gaps	69
Average number of gaps per sentence	2.3
Evaluations per gap-configuration	5.7
Number of evaluations by hint type: NONE	393
Number of evaluations by hint type: GoURMET	391
Number of evaluations by hint type: Google	389

Table 107: Summary of GF evaluation sr→en

5.2.7 Amharic

Status: Silver standard, completed

The statistics for the GF evaluation for am→en are shown in Table 108. The detailed boxplot of results is shown in Figure 31.

As may be seen, the boxes for Google and GoURMET clearly overlap, meaning that the difference in usefulness is not huge.

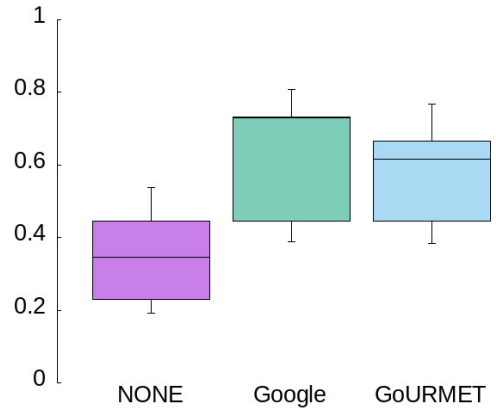


Figure 31: Results of GF evaluation am→en

Unique Evaluators	15
Number of unique gaps	74
Average number of gaps per sentence	2.5
Evaluations per gap-configuration	5
Number of evaluations by hint type: NONE	378
Number of evaluations by hint type: GoURMET	358
Number of evaluations by hint type: Google	374

Table 108: Summary of GF evaluation am→en

5.2.8 Kyrgyz

Status: Silver standard, completed

The statistics for the GF evaluation for ky→en are shown in Table 109. The detailed boxplot of results is shown in Figure 32.

The results for Kyrgyz GoURMET model into English remain below those achieved by Google.

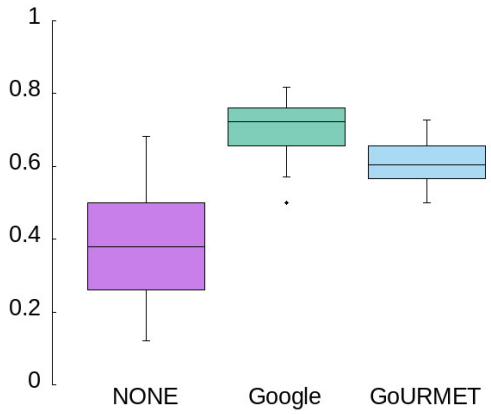


Figure 32: Results of GF evaluation ky→en

Unique Evaluators	16
Number of unique gaps	74
Average number of gaps per sentence	2.5
Evaluations per gap-configuration	5.4
Number of evaluations by hint type: NONE	400
Number of evaluations by hint type: GoURMET	403
Number of evaluations by hint type: Google	397

Table 109: Summary of GF evaluation ky→en

5.2.9 Macedonian

Status: Silver standard, completed

The statistics for the GF evaluation for mk→en are shown in Table 110. The detailed boxplot of results is shown in Figure 33.

Macedonian has been one of the highest performing GoURMET models, with a results range that is close to, but wider (i.e. with more variation) than Google's.

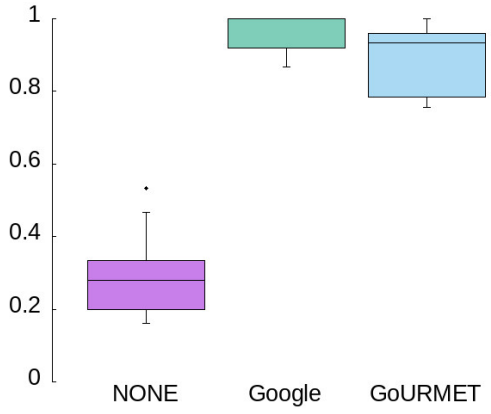


Figure 33: Results of GF evaluation mk→en

Unique Evaluators	15
Number of unique gaps	75
Average number of gaps per sentence	2.5
Evaluations per gap-configuration	5.1
Number of evaluations by hint type: NONE	351
Number of evaluations by hint type: GoURMET	417
Number of evaluations by hint type: Google	387

Table 110: Summary of GF evaluation mk→en

5.2.10 Hausa

Status: Silver standard, completed

The statistics for the GF evaluation for ha→en are shown in Table 111. The detailed boxplot of results is shown in Figure 34.

Although the overall performance level above the 0.6 range is on par or better than other GoURMET models, Google performs significantly better than GoURMET in this instance, with no overlap registered in the box charts.

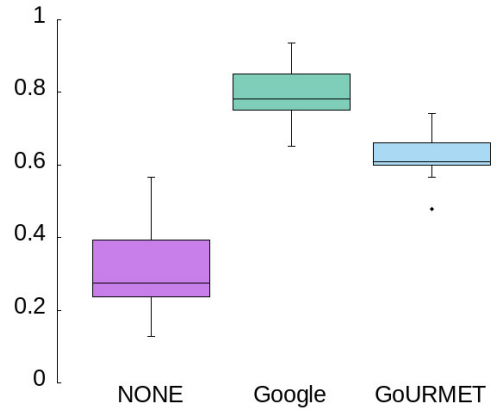


Figure 34: Results of GF evaluation ha→en

Unique Evaluators	16
Number of unique gaps	73
Average number of gaps per sentence	2.4
Evaluations per gap-configuration	5.4
Number of evaluations by hint type: NONE	398
Number of evaluations by hint type: GoURMET	404
Number of evaluations by hint type: Google	382

Table 111: Summary of GF evaluation ha→en

5.2.11 Igbo

Status: Silver standard, completed

The statistics for the GF evaluation for ig→en are shown in Table 112. The detailed boxplot of results is shown in Figure 35.

The boxplot for this language pair indicates that the overlap is limited to the maximum and minimum points, with a higher performance from Google. The findings correlate with automatic evaluation data.

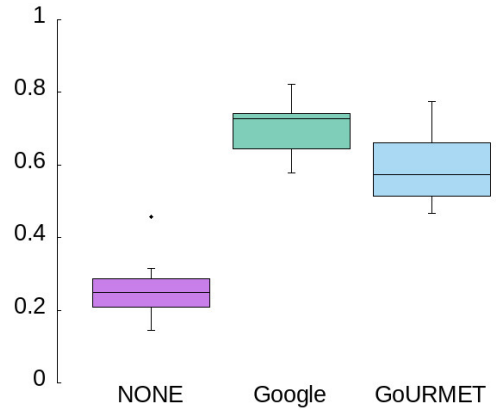


Figure 35: Results of GF evaluation ig→en

Unique Evaluators	14
Number of unique gaps	141
Average number of gaps per sentence	4.7
Evaluations per gap-configuration	4.7
Number of evaluations by hint type: NONE	665
Number of evaluations by hint type: GoURMET	648
Number of evaluations by hint type: Google	675

Table 112: Summary of GF evaluation ig→en

5.2.12 Tigrinya

Status: Silver standard, completed

The statistics for the GF evaluation for `ti→en` are shown in Table 113. The detailed boxplot of results is shown in Figure 36.

While this language pair is among the lowest scorers according to automated evaluations, the results of human evaluations are not significantly low, and GoURMET outperforms Google.

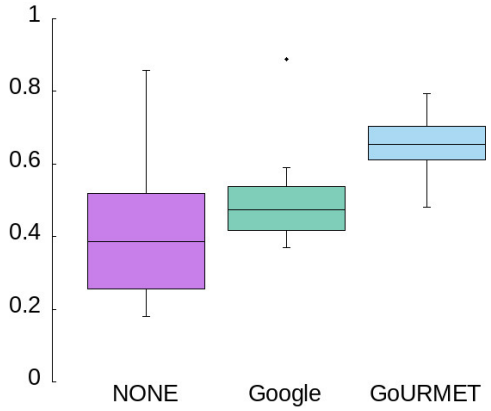


Figure 36: Results of GF evaluation `ti→en`

Unique Evaluators	12
Number of unique gaps	100
Average number of gaps per sentence	3.3
Evaluations per gap-configuration	4.0
Number of evaluations by hint type: NONE	400
Number of evaluations by hint type: GoURMET	416
Number of evaluations by hint type: Google	396

Table 113: Summary of GF evaluation `ti→en`

5.2.13 Pashto

Status: Silver standard, completed

The statistics for the GF evaluation for `ps→en` are shown in Table 114. The detailed boxplot of results is shown in Figure 37.

For this language pair Google appears to perform significantly better than GoURMET which appears to somewhat contradict the direct feedback we have received from the users in terms of accuracy.

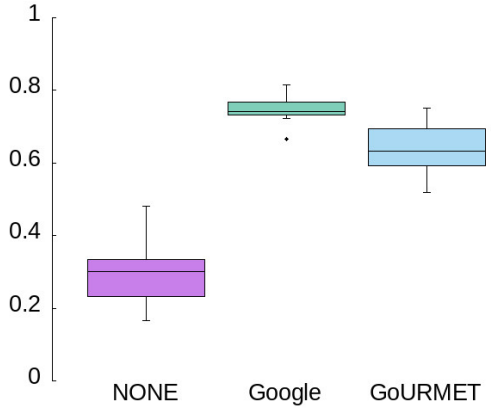


Figure 37: Results of GF evaluation `ps→en`

Unique Evaluators	17
Number of unique gaps	92
Average number of gaps per sentence	3.1
Evaluations per gap-configuration	6.1
Number of evaluations by hint type: NONE	567
Number of evaluations by hint type: GoURMET	555
Number of evaluations by hint type: Google	561

Table 114: Summary of GF evaluation `ps→en`

5.2.14 Burmese

Status: Silver standard, completed

The statistics for the GF evaluation for my→en are shown in Table 115. The detailed boxplot of results is shown in Figure 38.

Burmese is one of the two languages (the other being Tigrinya) where the median score GoURMET achieved into English is higher than Google. The boxes for Google and GoURMET overlap, meaning that the difference in usefulness is not significant. While the range is wider for GoURMET results, the median point as well as top scores are still marginally higher in comparison.

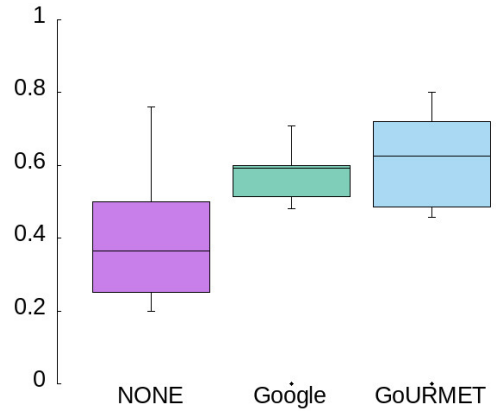


Figure 38: Results of GF evaluation my→en

Unique Evaluators	10
Number of unique gaps	83
Average number of gaps per sentence	2.8
Evaluations per gap-configuration	3.1
Number of evaluations by hint type: NONE	263
Number of evaluations by hint type: GoURMET	254
Number of evaluations by hint type: Google	245

Table 115: Summary of GF evaluation my→en

5.2.15 Yoruba

Status: Silver standard, completed

The statistics for the GF evaluation for yo→en are shown in Table 116. The detailed boxplot of results is shown in Figure 39.

Both Google and GoURMET overlap with the baseline (NONE) which indicates that neither Google nor GoURMET are too helpful for translating Yoruba into English. The results correlate with the other evaluation data reported in earlier sections and echo our experience throughout the process. Yoruba had proven to be a challenge in terms of compiling parallel data, and arriving at an adequate number of validated translations.

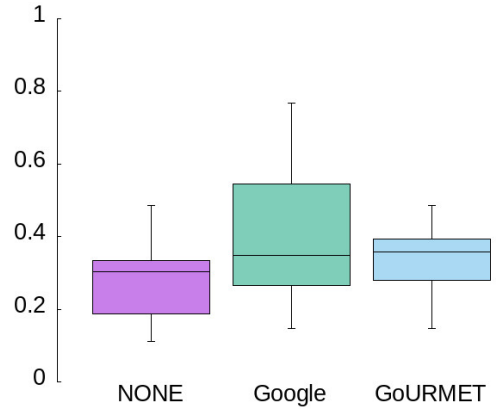


Figure 39: Results of GF evaluation yo→en

Unique Evaluators	10
Number of unique gaps	110
Average number of gaps per sentence	3.8
Evaluations per gap-configuration	3.1
Number of evaluations by hint type: NONE	348
Number of evaluations by hint type: GoURMET	336
Number of evaluations by hint type: Google	333

Table 116: Summary of GF evaluation yo→en

5.2.16 Urdu

Status: Silver standard, completed

The statistics for the GF evaluation for ur→en are shown in Table 117. The detailed boxplot of results is shown in Figure 40.

For this language pair both systems have been scored consistently, with Google placed slightly ahead of GoURMET.

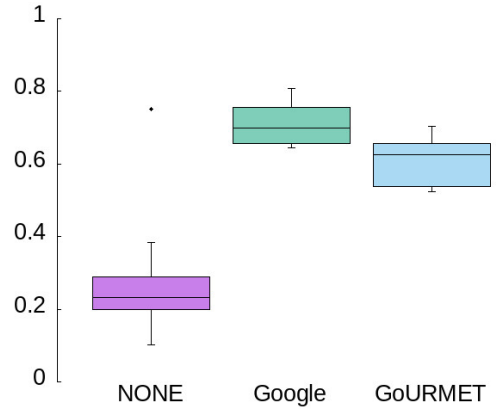


Figure 40: Results of GF evaluation ur→en

Unique Evaluators	12
Number of unique gaps	109
Average number of gaps per sentence	3.6
Evaluations per gap-configuration	3.9
Number of evaluations by hint type: NONE	421
Number of evaluations by hint type: GoURMET	431
Number of evaluations by hint type: Google	428

Table 117: Summary of GF evaluation ur→en

5.2.17 Turkish v2

Status: Silver standard, completed

The statistics for the GF evaluation for tr→en are shown in Table 118. The detailed boxplot of results is shown in Figure 41.

In the samples for the updated model, the boxes for Google and GoURMET clearly overlap, with both figures slightly higher that earlier results reported in D5.4. The median figures are also on par, suggesting the difference in usefulness is not significant.

It has to be noted however that the first tr→en GoURMET model was almost 15 BLEU points behind the Google model available at the time according to the interim report. Since then, Google's BLEU scores for tr→en have also improved by 5 points as demonstrated in Figure 7 in this document. Therefore, the leap achieved by redeveloping the Turkish model is remarkable.

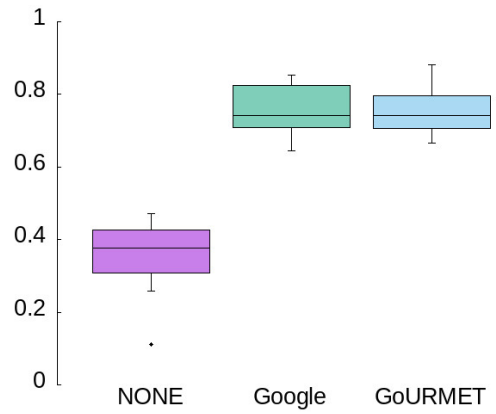


Figure 41: Results of GF evaluation tr→en v2

Unique Evaluators	12
Number of unique gaps	92
Average number of gaps per sentence	3.1
Evaluations per gap-configuration	4.3
Number of evaluations by hint type: NONE	392
Number of evaluations by hint type: GoURMET	396
Number of evaluations by hint type: Google	410

Table 118: Summary of GF evaluation tr→en v2

5.2.18 Gap Filling Evaluation Findings

All of the languages developed during months 19 to 42 have been evaluated to a 'silver' standard with regards to the gap filling tasks. The highest scores were achieved by **Serbian** and **Macedonian**, with the GoURMET median figures in boxplots around or above the 0.8 mark (80% gap-filling success rate).

Bulgarian, **Tamil**, **Amharic**, **Kyrgyz**, **Macedonian**, **Hausa**, **Tigrinya**, **Pashto**, **Burmese**, **Yoruba**, **Urdu** and **Turkish v2** achieved median figures of over 0.6. This indicates that the machine translation provided the correct word for each gap in the reference sentence, for the annotator to then complete it correctly, in more than 60% of cases).

Where MT helps, success rates should clearly distinguish from the no-hint success rates, and this happens with many systems in our data. All of the cases demonstrate the hints provided by GoURMET were helpful compared to non-hint scenarios. The difference was most pronounced for **Serbian**, **Macedonian** and **Pashto** and least pronounced for **Gujarati**, **Tigrinya**, **Burmese** and **Yoruba**.

Compared with the highest scorers for automated evaluation results (e.g. spBLEU, COMET) of translations in to English, the languages that were also rated highly by human evaluators were **Macedonian**, **Urdu** and **Turkish**.

It should be noted that all the gap-filling sentences were derived from the news domain. However, evaluators selected for their skills in English would not necessarily command knowledge of references to the people and events in the test sentences from that particular language or country.

With further time, it would be interesting to take a closer look at the extent to which an evaluator's background knowledge (or lack of it) about the stories had any bearing on the scores achieved for non-hint stories.

6 Results of Post-Edit Evaluation and Benchmarking

6.1 Results of Post-Edit Evaluation

We conducted 'gold' standard post-edit evaluation on translations from English into and from Urdu, Serbian and Turkish.

As explained in section 2.3, post-editing effort can be represented by the number of keystrokes or the number of deletions, insertions and substitutions made. In this research, we have used a quantitative measure of technical effort called Translation Error Rate (TER) Snover et al. (2006), which reflects the number of editing operations necessary to transform the MT output into the final version. Technical effort is more related to cognitive effort than to post-editing time, because post-editing time is strongly dependent on sentence length Popović et al. (2014). In general, TER is calculated as a number of changed words in the sentence divided by the total number of words.

TER edits include insertions, deletions, substitutions and shifts and are calculated over whole words, so if the annotator only changes the ending of a word, TER still considers the whole word as changed. All edits count as 1 edit. Shift moves a sequence of words within the hypothesis, and a shift of any sequence of words (any distance) is only 1 edit. When the TER score is under 30%

it is generally considered to be easier to post-edit than to translate from scratch [4], which is more or less the same as editing three words out of ten.

There were nine articles edited for Urdu, nine articles for Serbian and ten articles for Turkish. In total, 19 language combinations were edited.

6.1.1 Post-Edit Score Summary

You can see our results in Table 119. The TER score was calculated with the Sacrebleu script Post (2018b) with the standard settings: lowercase and including punctuation.

	en-ur	ur-en	en-sr	sr-en [a]	en-tr	tr-en
No. stories	4	5	2	7	6	4
No. words MT	1414	147	1258	6312	719	649
No. words PE	1432	152	1126	6326	713	652
TER score	20.0	10.2	63.9	0.8	36.9	2.1

Table 119: Post-editing effort as captured by the TER score (lower is better), including statistics about the test sets.

[a] The Serbian to English post-edit results were deleted due to a technical error. However, feedback from the participants is reported in section 6.1.4

The TER score into English is generally very low, which is highly encouraging, suggesting that few post-edits are required for using the translations. The out of English scores are higher:

The en–ur TER is still quite low at 20.0%.

The en–sr TER score is very high. This partially reflects the quality of the machine translation output, but also reflects the difficulty of using highly trained journalists to do minimal edits to a translation.

The en-tr TER score is also high. Turkish is a very morphologically rich language, and even small morphological changes would result in a TER penalty.

It should be noted that post-editing translations is not journalists' normal mode of operation, and they are likely to edit the output heavily to reflect the interests of their readers. It is therefore important to take the editors' preferences into account (see sections 6.1.3 to 6.1.5 below for more detailed commentary).

A gloss that needs to be edited can still be useful to speed up content creation in many languages, even if the TER score is over 30, especially as this is not a standard translation use case, but a content creation use case.

6.1.2 Post-Edit Feedback

As introduced in section 2.3, in addition to us using their edits to calculate the TER scores, participants were asked to provide feedback on the accuracy of the translation and probed on what aspect of post-edits took them the longest time.

[4] https://kantanmtblog.com/2015/07/28/what-is-translation-error-rate-ter/

Participants were asked:

1. **Accuracy** – On a scale of 1-10, the 'accuracy' of the translation (not the language quality or style) was...
    - 1 – very poor
    - 10 – amazing
2. **Quality** – On a scale of 1-10, the 'quality' of the translation (good construction, natural flow, style) was...
    - 1 – very poor
    - 10 – amazing
3. **Editing time** – The time editing this article took was...
    - 1 – too long to be useful
    - 5 – very short
4. **Usefulness** – Having a basic translation to start from was... (please select as applicable)
    - Confusing / misleading
    - Helpful / time saving
    - Great / smooth
    - Other...
5. **Took longest** – What took the longest time was...
    - Checking facts in my language against the original
    - Fixing sentences
    - Improving style
    - Other...

Most answers (22 of 28) rated the accuracy of the translations between 8-10. Having MT to start reversioning process was considered time-saving (18) and great/smooth (7). In 3/28 cases, starting with the machine translated text was considered misleading/confusing.

The feedback on what took the longest time is summarised in Figure 42.

It has to be noted, however, that in dealing with translations into target language, the evaluators could not speak the source languages and had to rely on the English translations for accuracy.

6.1.3 Results for Urdu

Among the three languages covered, Urdu had the lowest scores. The evaluator's scores were lower into English and higher into Urdu, which correlated with the time it took him to post edits and the perceived helpfulness of using MT (see Tables 120 and 121).

In several cases, the issues raised included the confusion of third person genders, inconsistency in spellings, and issues with names and named entities.

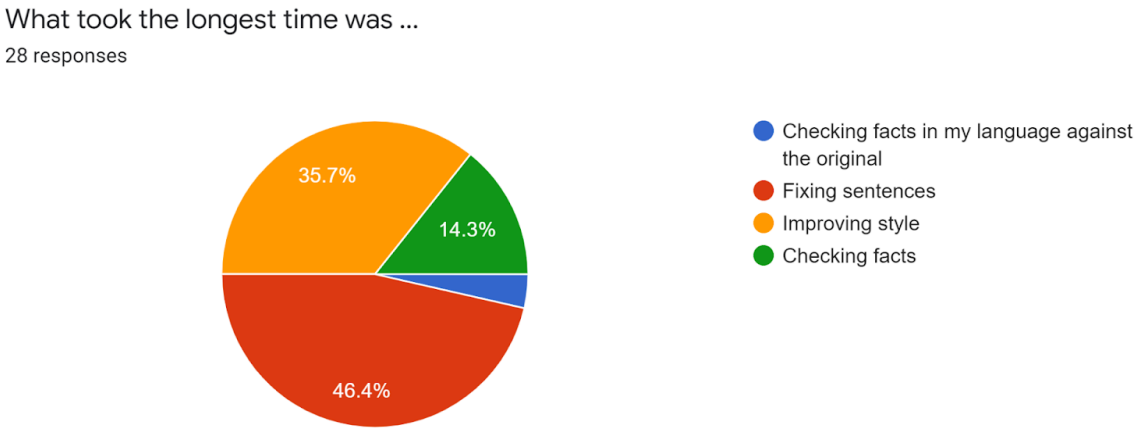


Figure 42: Chart showing what took the longest time during post-editing

Language pair	Accuracy	Quality	Editing time	Usefulness
	6	5	3	Confusing / misleading
	8	7	4	Helpful / time saving
	7	6	3	Confusing / misleading
	6	5	3	Helpful / time saving

Table 120: Evaluators' scores for editing text translated from Urdu into English

Language pair	Accuracy	Quality	Editing time	Usefulness
	9	8	5	Great / smooth
	8	6	5	Great / smooth
	10	8	5	Great / smooth
	8	6	3	Helpful / time saving

Table 121: Evaluators' scores for editing text translated from English into Urdu

In one story, 'the name of [the] Prime Minister was completely wrong in nearly all instances except once'. The evaluator also noted for another article that the name of the main protagonist was again translated instead of being kept intact as a named entity, adding that 'the name had to be fixed in almost EVERY instance except once, and I could not understand why it was able to translate the name correctly in one instance but in every other, it failed'.

6.1.4 Results for Serbian

The Serbian team, who publish both in Latin and Cyrillic alphabets, worked with nine articles in the range of 1500-2000 words in length. While their scores were predominantly between 8 and 9 for quality, the only exception to this was when the source material came from Hindi, a comparatively lower- resourced language (see Tables 122 and 123).

Language pair	Accuracy	Quality	Editing time	Usefulness
Serbian to English	8	8	4	Helpful / time saving
Serbian to English	9	7	3	Helpful / time saving
Serbian to English	8	8	4	Helpful / time saving
Serbian to English	8	7	4	Helpful / time saving

Table 122: Evaluators' scores for editing text translated from Serbian into English

Language pair	Accuracy	Quality	Editing time	Usefulness	Took longest
Spanish to Serbian	9	7	4	Helpful / time saving	Improving style
Russian to Serbian	8	9	3	Helpful / time saving	Fixing sentences
Spanish to Serbian	9	8	5	Helpful / time saving	Improving style
Spanish to Serbian	9	8	4	Helpful / time saving	Improving style
Hindu to Serbian	8	6	3	Helpful / time saving	Checking facts

Table 123: Evaluators' scores for editing text translated from English into Serbian

The Serbian Service did not directly comment on the edits required on the machine translated texts on a case-by-case basis. However, they did comment that the time spent checking stories was definitely shorter than translating them from scratch.

Due to social and cultural affinities, most of the content was selected from Russian and Spanish sources.

6.1.5 Results for Turkish

The sample included six articles from various languages, translated first into English and then to Turkish. The source languages were both from GoURMET languages (Tamil, Turkish) and non-GoURMET languages (French, Spanish). There were also four articles translated from Turkish into English.

Generally, perceived quality and usefulness scores were higher when translating into English (see Tables 124 and 125).

Language pair	Accuracy	Quality	Editing time	Usefulness	Took longest
Turkish to English	10	9	5	Great / smooth	Improving style
Turkish to English	8	8	4	Great / smooth	Improving style
Turkish to English	10	9	5	Great / smooth	Improving style
Turkish to English	7	7	3	Helpful / time saving	Improving style

Table 124: Evaluators' scores for editing text translated from Turkish into English

Language pair	Accuracy	Quality	Editing time	Usefulness	Took longest
French to Turkish	9	9	5	Great / smooth	Improving style
French to Turkish	8	6	3	Helpful / time saving	Fixing sentences
Portuguese to Turkish	7	5	3	Helpful / time saving	Fixing sentences
Spanish to Turkish	9	7	3	Helpful / time saving	Fixing sentences
Gujarati Turkish	9	7	4	Helpful / time saving	Fixing sentences
Tamil to Turkish	3	2	1	Confusing / misleading	Checking facts in my language against the original

Table 125: Evaluators' scores for editing text translated from English into Turkish

When the perceived quality was higher, edits were limited to improving style. However, it has proven to be a challenge for lower-resourced languages such as Tamil, where the facts needed to be checked extensively, and the language sounded substandard (e.g. in one article translated from Gujarati, as tester commented that 'it is almost like the article has been written by someone in Gujarat who did not know English very well').

An illustration of the point above is the sentence:

'They have married 182 senior citizens so far and arranged 12 couples in a live-in relationship.'

Which might appear in print along these lines:

'They have assisted 182 golden agers to get married so far and matched 12 couples who are living together.'

Although the construction in Turkish appeared good overall, the end user needed to be on their guard for checks, since there was at least one incident per article of 1000-1500 words where words were replaced with their antonyms in translated versions, entirely altering the meaning. There was one case of a hallucination, where a whole sentence was generated from scratch.

For instance, the word *'unpasteurized'* was translated as *'pasteurized'* and *'enthusiasm'* as *'disgust'*. Particularly in the Tamil to Turkish example, there were significant mistranslations, such as *'coronavirus'* being translated as the *'Qoran virus'* in a story relating to interreligious tensions, which might prove highly inflammatory.

In a similar vein, while quoting a US foreign affairs spokesman, one sentence read:

'Nothing about Ukraine should be Ukrainian.'

When what he said was:

'Nothing about Ukraine should be decided without Ukrainians.'

This final example illustrates the crux of editorial concerns around risk of reputational damage from MT. The output in this case is not just wrong but highly problematic, politically. Had it been published as translated, it could potentially have caused serious issues for the BBC that would have resulted in apologies and retractions.

6.1.6 Conclusions from Post-Editing

Despite the challenges of recruiting annotators from the BBC World Service in the final stages of the project, the post-editing exercise has allowed us to gauge the perceived usefulness of machine translated content in a real-life setting. Unlike automated evaluations, which primarily focus on word-level, and direct assessment, which focuses on sentence level, this exercise has involved entire articles of up to 2000 words and allowed us to observe the range and consistency of output.

For instance, consistency across named entities appeared to suggest an area for improvement. In one Turkish sample, the organisation Tablighi Jamaat appeared as *Dublin Jamaat*, *Tabloid Jamaat* or *Taplek Jamaat*, respectively echoing the feedback from an Urdu annotator on named entities.

There were two difficulties BBC has faced in the process:

Firstly, there was an issue with not having sufficient time and access to the right range of languages for the exercise. In an ideal world, it would be helpful to be able to include a wider selection of languages and focus on certain subsets for post-edits, such as Kyrgyz–Russian, Turkish–Kyrgyz, Tamil–Hindi, Gujarati–Tamil, etc. Operational challenges such as the Russia-Ukraine War and business needs (resourcing) have hindered setting up a more comprehensive framework.

Secondly, the exercise could have benefited from more precise sets of goals and metrics set out in advance. Agreeing on what factors to focus on from a research perspective across the Consortium, could have guided the BBC better in making a case for the task and setting up the structures to elicit useful data. Despite the reluctance of editorial leaders to include it in this instance, being able to measure time spent for post-editing in such exercises remains key to draw more effective comparisons.

Machine translation clearly speeded up translation and reversioning processes and is viewed generally positively by users for the selected set of languages. It is particularly useful to spot trends and interesting content across the range of services/providers, and more widely for monitoring and gisting purposes.

The differences in scores suggest annotators were more rigorous when editing the content with a view to publish it in front of real audience members. However, in cases of translating into English, the output text would often serve as a draft for internal consumption to promote a particular story to other teams, which would in turn be translated into another language, so stylistic considerations were seemingly less prominent and the post-editing effort appears to have been lighter.

However, the findings also clearly suggest the need to have a competent human post-editor to mitigate mistranslations. This is not a fail-safe process in a fast-paced newsroom where the post-editor might be required to deal with a text originating in a language they do not speak. Where a sentence appears ‘obviously wrong’, issues are easier to detect and fix. In cases where errors are masked by competent flow of language, they might easily go unnoticed. Therefore, users would need assistive infrastructures such as translation quality estimations to have potential mistranslations flagged, to be swiftly validated and corrected in communication with the originators. This is an area that BBC News Labs teams aim to explore further in the future as part of a validation workflow.

At the point of wider roll-outs of machine translated content, it would also be advisable to run events with editorial team members to raise awareness about improving MT output by being mindful about the language in the source texts (e.g. writing in shorter sentences, avoiding double negatives or long noun phrases). Last but not least, developing competencies around named entity recognition and consistency, potentially by exploring translation memories, might be another way to enhance the post-editing experience.

6.2 Results of DW Benchmarking

DW's benchmarking is based on four types of evaluation:

- Automated back translation in case no reference text is available
- Automated evaluation using BLEU scores by means of a reference translation
- Human evaluation by comparing MT output from different engines
- Human evaluation by means of post-editing

We have five standard texts (video manuscripts) which we use to evaluate MT output in different languages. In addition, ad hoc texts are used for specific languages.

We ran a back translation on all five manuscripts for all the GoURMET language pairs, which gives us an initial indication of the quality output.

From this, we decided to focus on those DW languages with the highest scores: Bulgarian, Macedonian, Serbian, Pashto and Turkish for more in-depth human evaluation. Our focus is translation from English into those languages. We produced at least one reference translation for each selected language in order to proceed to the next stages. We ran the texts through four engines (i.e. GoURMET, Google, Azure, Facebook (Easy NMT)), and subsequently ran it through the Tilde BLEU scoring tool to get an automated ranking.

Language	GoURMET	Google	Azure	Facebook
Bulgarian	38.27	38.06	33.07	33.07
Macedonian	58.06	53.99	48.07	41.70
Pashto	30.78	33.46	20.58	11.83
Serbian	55.01	45.75	44.08	34.72
Turkish	21.10	29.78	28.97	20.69

Table 126: Comparative automated BLEU scores

We also asked the editors to post-edit one or more translations from English and judge the effort that is required for this work. In addition, they were asked to compare and assess output from different machine translation engines, post-editing shows how much of the text has been edited and how much could be retained as such.

Post-editing is DW's primary goal for the use of machine translation and is becoming standard practice. The smaller language departments in particular have limited resources and a large part of their content consists of translations from English – and to some extent also German – source

material. Hence DW's major ongoing project to roll out its plain X HLT tool for automation-supported translation, subtitling and voiceover.

The opinion of the editors in terms of post-editing and overall performance of the GoURMET MT can be summarised as follows where evaluation focused on the translation from English into the five selected target languages:

Bulgarian 'It becomes immediately obvious that the text has been translated by a machine. It's hard to post-edit this. DeepL provides a much better basis.'

Macedonian 'Very good translation.'

Pashto 'My overall impression about GoURMET's translation is that it is not bad. It is comprehensive enough to understand the content. It conveys the real message of the text in most cases and it can be worked with. But of course the translation would need editing afterwards. It can be published but only after editing it. Good job! Really. Some problems do exist with gender markers of Pashto language. Sentences like "My wife is a doctor herself." contain words like "wife" and "herself" which can help the machine translate the sentence into Pashto with the correct gender, but it doesn't. Also, there are many words (especially technical terms) that cannot yet be translated into Pashto because such words do not yet exist in the Pashto language. Therefore, it can be helpful to leave such words untranslated in English alphabets between the Pashto translation or convert them to a Pashto transcription.'

Serbian 'Great starting point for editing. Please note: in Serbian we do not write foreign names in original but as we speak them. There is a set of rules for that. For instance, last name Torrado would be simply Torado in Serbian.'

Turkish Editor 1: 'Overall it's a perfect translation. There is one major flow though. The misgendering issue could be improved. In Turkish, pronouns are gender neutral. In the original text the person Francisco Torrado is a man. But in the translation, the tool assumes that he's a woman and then misgenders him.' Editor 2: 'One of the main issue is the tenses. In Turkish, present continuous tenses are preferred in some occasions, despite the English version using just present tense.'

The main results are summarised in Tables 127 to 129. The overall user rating for the GoURMET models compared to some other engines are shown in Table 130. In each case 1 = very poor, 2 = poor, 3 = acceptable, 4 = good, and 5 = very good.

	Bulgarian	**Macedonian**	**Pashto**	**Serbian**	**Turkish**
Accuracy	4	5	3	5	4
Capitalisation	5	5	-	5	5
Punctuation	5	5	3	5	5

Table 127: DW benchmarking results at word level (1 = very poor ↔ 5 = very good)

	Bulgarian	**Macedonian**	**Pashto**	**Serbian**	**Turkish**
Accuracy	4	5	2	4	4
Capitalisation	5	5	4	5	5
Fluency	3	5	3	4	4
Completeness	4	5	4	4	4
Punctuation	5	5	4	5	5

Table 128: DW benchmarking results at sentence level (1 = very poor ↔ 5 = very good)

	Bulgarian	**Macedonian**	**Pashto**	**Serbian**	**Turkish**
Accuracy	4	5	3	4	5
Capitalisation	5	5	4	5	5
Fluency	3	5	2	4	5
Completeness	4	5	4	4	5
Punctuation	5	5	4	5	5

Table 129: DW benchmarking results at document level (1 = very poor ↔ 5 = very good)

Target language	**GoURMET**	**Google**	**Azure**	**Facebook**
Bulgarian	4.00	4.47	4.40	4.60
Macedonian	5.00	5.00	4.38	2.77
Pashto	4.00	4.53	4.40	4.60
Serbian	4.92	4.46	4.20	4.30
Turkish	4.28	4.50	4.69	3.46

Table 130: DW benchmarking results overall user rating (1 = very poor ↔ 5 = very good)

Of course, these numbers are subjective, as these are human evaluations and different people have different standards. However, the comparison between the engines per language is consistent per evaluator, and we get an idea of the usefulness, especially in combination with the comments given by the editors.

In terms of metric, comparing post-editing results for these languages gives us an indication of what percentage of MT text could be retained (see Table 131).

Target Language	**Score - Recovered text**
Bulgarian	38.27
Macedonian	58.06
Pashto	52.78
Serbian	82.46
Turkish	30.53

Table 131: DW benchmarking post-editing results

From this user testing, complemented by BLEU scores, we can conclude that some of the GoURMET models are indeed usable in media production environments. There are some tools that perform better (e.g. Google outranks GoURMET in three of these languages but the difference is not that big). Thus, they are well suited for monitoring or comprehension purposes.

In addition, using them for content creation, bringing it to publishable quality, some effort is required, but overall, it is still considered doable by the editors. Such models are therefore valid options for monitoring or digestion, and as alternative engines in case of required cost reduction or enhanced control. In particular in case sending content to third-party translation tools in the cloud may not be allowed, due to the sensitivity of the data.

7 Conclusions

As the coordinator of D5.6, the BBC's goal was delivering 'a complete set of results for both automated and human evaluation' as conducted by the media partners BBC and DW. From that perspective, GoURMET has succeeded in compiling and evaluating a wide range of samples derived from under-resourced languages in the news domain.

The project had the highly ambitious goal of building quality MT models for media production environments. Our evaluations across the board indicate that the project team have succeeded in developing useful models that in most cases can go head-to-head with technology giants such as Google, and surpassing their output in some cases (i.e. Burmese and Tigrinya). More than anything else, this has proven the viability of being able to develop successful specialist models for the domain.

The findings of the automated evaluation and human evaluation work, both for direct assessment and gap filling exercises, broadly correlate. The post-edit results for the selected languages validate the trends we have seen from the other types of evaluation. Across all languages, Macedonian, Bulgarian, Serbian provide the best outcomes while Turkish, Hausa, Amharic and Tamil are also promising.

With the high expectations of quality and accuracy in global newsrooms, it is likely to be some time before any machine translation tool can provide automated, unmediated content creation that is acceptable from a journalistic point of view. The translation models are not yet ready to be used in this context, although Serbian and Macedonian are extremely close. However, all the GoURMET models can add value to the newsroom production workflow, providing a previously unavailable overview of content for gisting and monitoring purposes.

References

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301. URL https://www.aclweb.org/anthology/W19-5301.

Susan Bassnett and Esperanca Bielsa. *Translation in Global News*. Taylor & Francis Group, September 2008.

Ondrej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W18-6401.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/W07-0718.

M Esplà-Gomis and M.L. Forcada. Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93:77–86, 2010.

Ignatius Ezeani, Paul Rayson, Ikechukwu E. Onyenwe, Chinedu Uchechukwu, and Mark Hepple. Igbo-english machine translation: An evaluation benchmark. *CoRR*, abs/2004.00648, 2020. URL https://arxiv.org/abs/2004.00648.

Marina Fomicheva and Lucia Specia. Taking MT evaluation metrics to extremes: Beyond correlation with human judgments. *Computational Linguistics*, 45(3):515–558, September 2019. doi: 10.1162/coli_a_00356. URL https://aclanthology.org/J19-3004.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.

Barry Haddow, Rachel Bawden, Antonio Valerio Micheli Barone, Jindřic Helcl, and Alexandra Birch. Survey of low-resource machine translation. *Computational Linguistics*, 2022.

Geoffrey S Koby. Post-editing of machine translation. *The Encyclopedia of Applied Linguistics*, 2012.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation, 2021. URL https://arxiv.org/abs/2107.10821.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 11 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00343. URL https://doi.org/10.1162/tacl_a_00343.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5302. URL https://aclanthology.org/W19-5302.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.21. URL https://aclanthology.org/2021.acl-long.21.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA, July 2002.

Maja Popović. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, 2015.

Maja Popović, Arle Lommel, Aljoscha Burchardt, Eleftherios Avramidis, and Hans Uszkoreit. Relations between different types of post-editing operations, cognitive effort and temporal effort. In *Proceedings of the 17th annual conference of the european association for machine translation*, pages 191–198, 2014.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6319.

Matt Post. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*, 2018b.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 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, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6488. URL https://aclanthology.org/W18-6488.

Andrew Secker, Susie Coleman, Mikel L. Forcada, Anna Blaziak, Rachel Bawden, Radina Dobreva Felipe Sánchez-Martínez, and Víctor M. Sánchez-Cartagena. GoURMET: Deliverable 5.3 - initial integration report, 2020.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 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, 2006.

Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem, Tewodros Abebe, Wondimagegnhue Tsegaye, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. Parallel corpora for bi-directional statistical machine translation for seven Ethiopian language pairs. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 83–90, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-3812.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 205–215, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.19.

Jian Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan, Xia Song, and Furu Wei. Multilingual machine translation systems from Microsoft for WMT21 shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 446–455, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.54.

Anna Zaretskaya, Mihaela Vela, Gloria Corpas Pastor, and Miriam Seghiri. Measuring post-editing time and effort for different types of machine translation errors. *New Voices in Translation Studies*, 15:63–91, 2016.

Appendix A Gap-Filling Results Combined

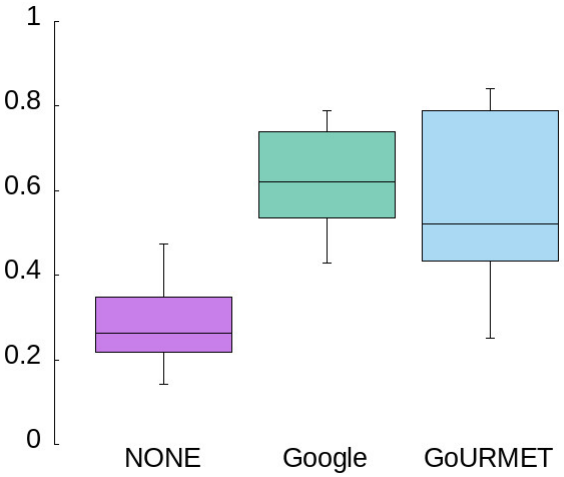
This appendix shows the gap-filling results combined without commentary, so they can be viewed at a glance:

The statistics for the GF evaluation for all languages into English are shown in Table 132.

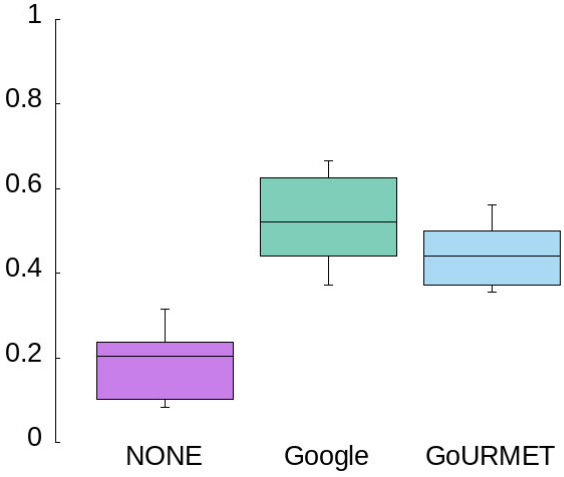
The box plots from Figures 25-41 are reproduced on page 81.

GoURMET language ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
ISO 639-1 language code	sw	gu	tr	bg	ta	sr	am	ky	mk	ha	ig	ti	ps	my	yo	ur	tr
Unique Evaluators	18	15	18	19	13	17	15	16	15	16	14	12	17	10	10	12	12
Number of unique gaps	70	143	83	72	95	69	74	74	75	73	141	100	92	83	110	109	92
Average number of gaps per sentence	2.33	4.77	2.80	2.40	3.17	2.30	2.47	2.47	2.50	2.43	4.70	3.33	3.07	2.8	3.8	3.63	3.07
Evaluations per gap-configuration	6.33	5.07	6.00	6.33	4.38	5.67	5.00	5.41	5.13	5.41	4.70	4.04	6.10	3.1	3.1	3.91	4.34
Number of evaluations by hint type:																	
NONE	448	725	498	457	425	393	378	400	351	398	665	400	567	263	348	421	392
GoURMET	439	725	498	456	412	391	358	403	417	404	648	416	555	254	336	431	396
Google	443	725	498	455	411	389	374	397	387	382	675	396	561	245	333	428	410

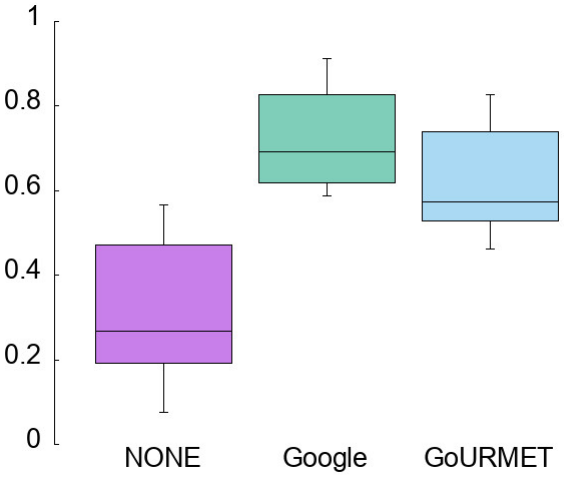
Table 132: Summary of GF evaluation for all languages into English



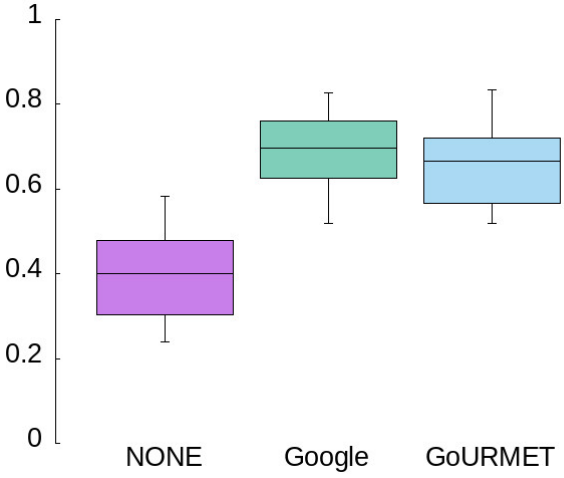
1. Swahili sw→en



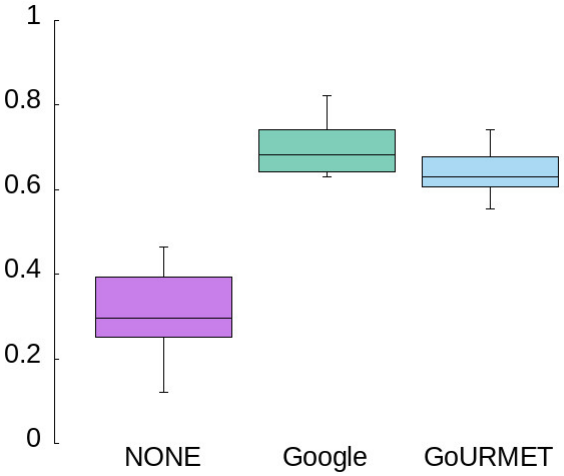
2. Gujarati gu→en



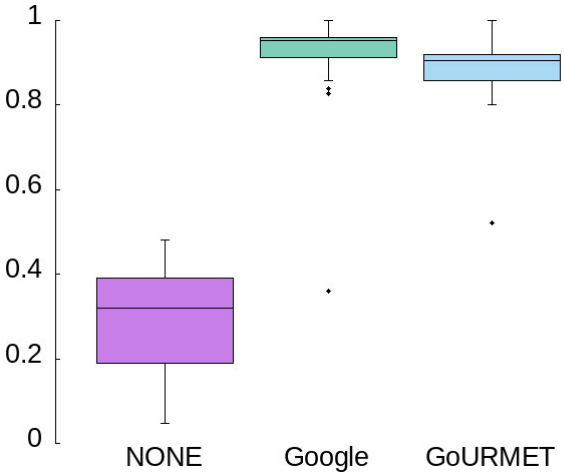
3. Turkish tr→en v1



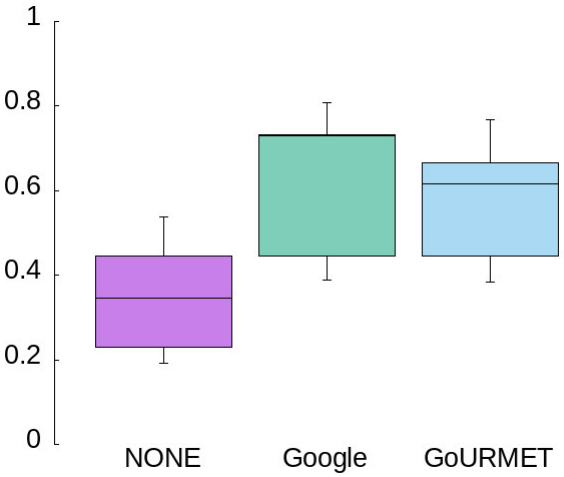
4. Bulgarian bg→en



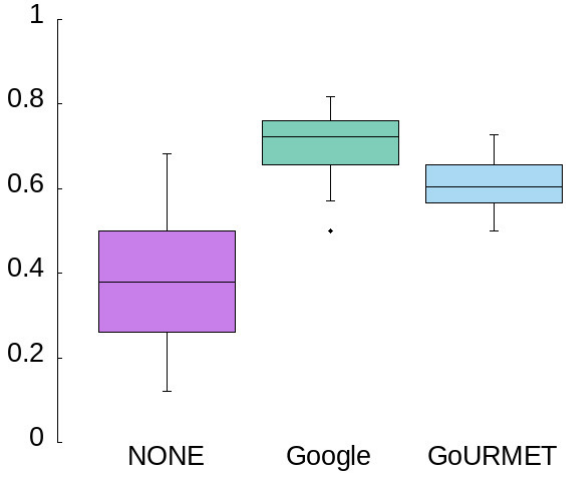
5. Tamil ta→en



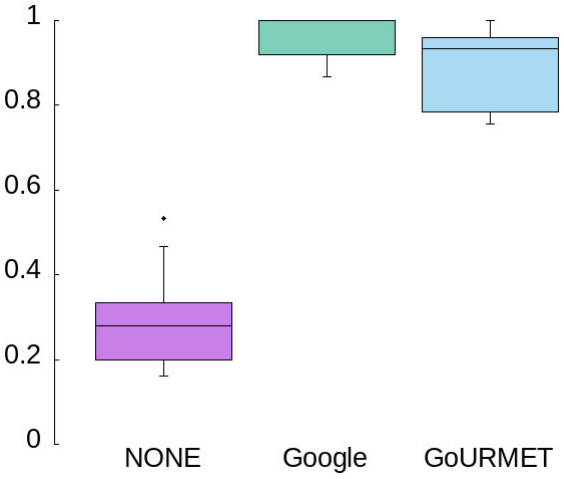
6. Serbian sr→en



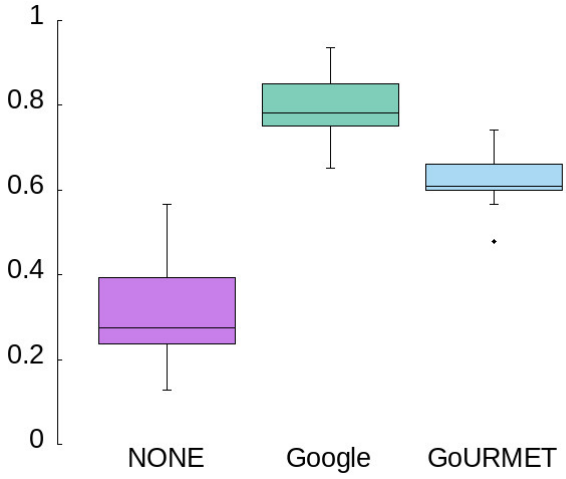
7. Amharic `am`→`en`



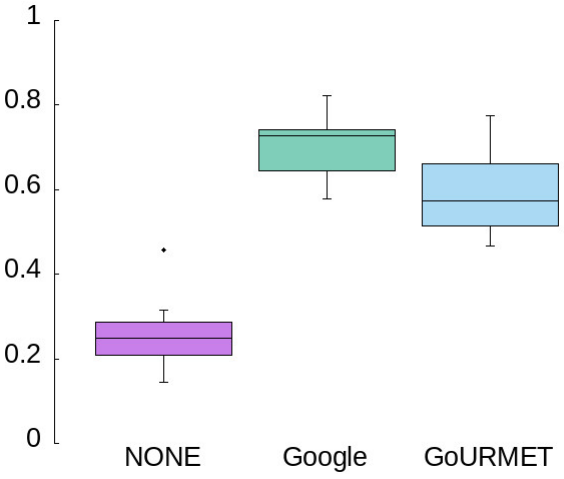
8. Kyrgyz `ky`→`en`



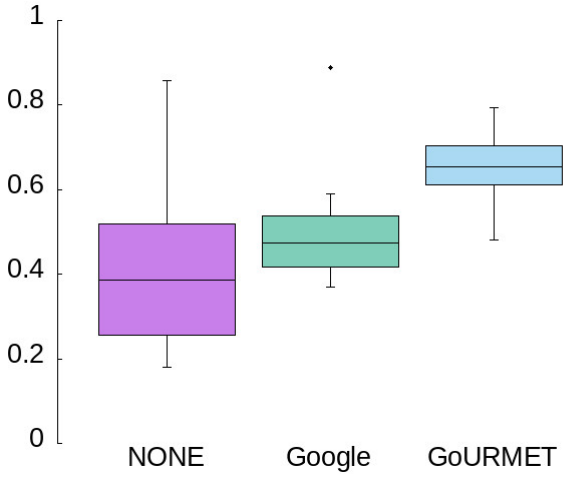
9. Macedonian `mk`→`en`



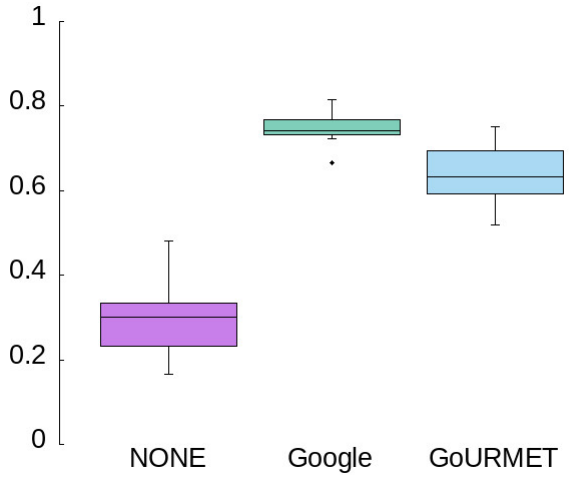
10. Hausa `ha`→`en`



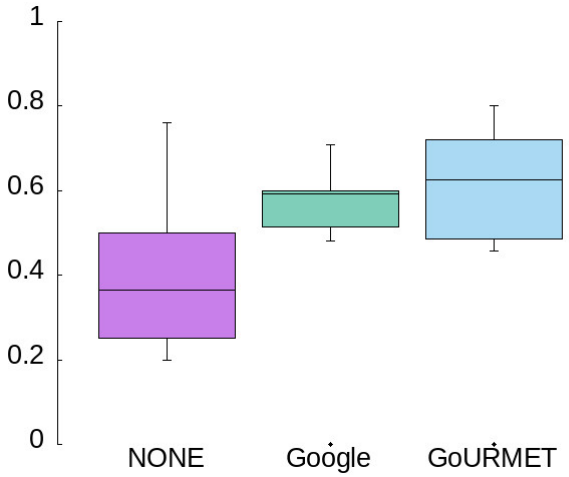
11. Igbo `ig`→`en`



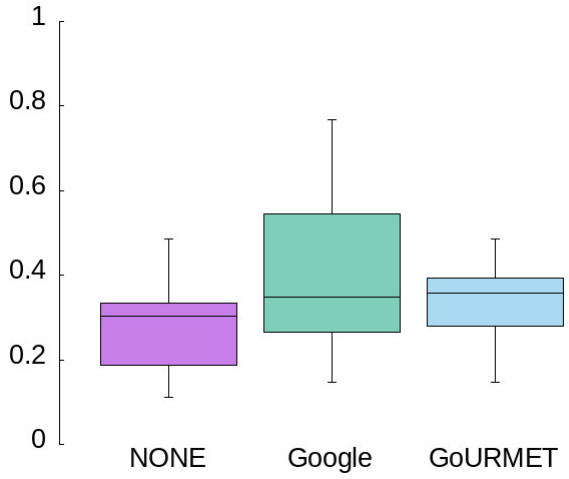
12. Tigrinya `ti`→`en`



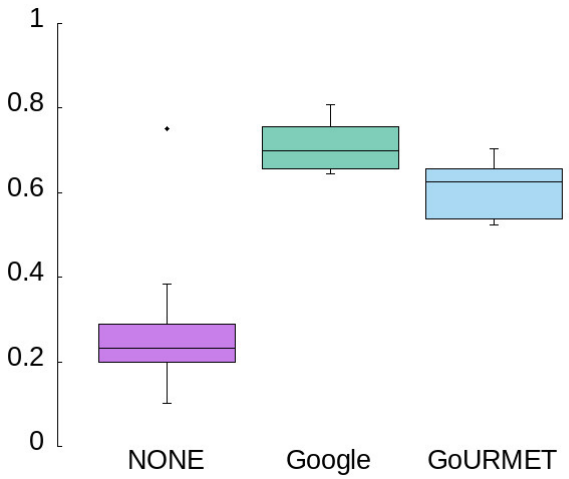
13. Pashto ps→en



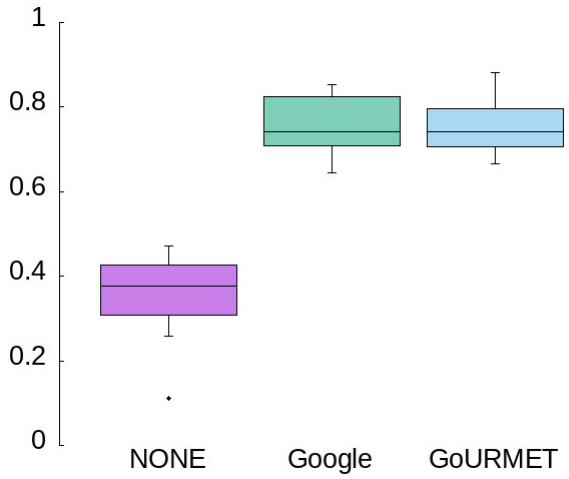
14. Burmese my→en



15. Yoruba yo→en



16. Urdu ur→en



17. Turkish tr→en v2

ENDPAGE

GoURMET

H2020-ICT-2018-2 825299

D5.6 GoURMET Final progress report on evaluation