Findings of the WMT'22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages

David Ifeoluwa Adelani♠ Md Mahfuz Ibn Alam† Antonios Anastasopoulos† Akshita Bhagia◊
Marta R. Costa-jussà♡ Jesse Dodge◊ Fahim Faisal† Natalia Fedorova‡ Christian Federmann⊗
Francisco Guzmán♡ Sergey Koshelev‡ Jean Maillard♡ Vukosi Marivate# Jonathan Mbuya†
Alexandre Mourachko♡ Safiyyah Saleem♡ Holger Schwenk♡ Guillaume Wenzek♡
†George Mason University ♠University College London ♡Meta AI
⊗Microsoft Research ‡Toloka #University of Pretoria ◊AI2

Abstract

We present the results of the WMT'22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages. The shared task included both a data and a systems track, along with additional innovations, such as a focus on African languages and extensive human evaluation of submitted systems. We received 14 system submissions from 8 teams, as well as 6 data track contributions. We report a large progress in the quality of translation for African languages since the last iteration of this shared task: there is an increase of about 7.5 BLEU points across 72 language pairs, and the average BLEU scores went from 15.09 to 22.60.

1 Introduction

A large portion of the world's population speak *low-resource languages*, and would benefit from improvements in translation quality on their native languages. Recent advances in translation quality, particularly from massively multilingual models (Fan et al.; Ma et al., 2021), have enabled progress in the translation quality of low-resource languages. However, many languages have seen little to no progress. This is particularly true for African languages. For example, in the 2021 Large Scale Multilingual Evaluation shared task (Wenzek et al., 2021), there was a progress[2] of +19.3 avg. BLEU when translating into languages like Irish and Welsh (included in the Other Indo-European grouping), but only a progress of +3.5 avg. BLEU when translating into languages like Fula and Igbo (included in the Nilotic/Atlantic Congo grouping).

The African continent is home to a rich diversity of languages. Around a third of the world's living languages are from Africa and only a small fraction of the resources in NLP and Machine translation are dedicated to them (Orife et al., 2020). As a result, African language speakers do not benefit from language technologies similarly to other Global North language speakers (Blasi et al., 2022). A major (but not the sole) hurdle to building language technologies for African languages is data availability (Joshi et al., 2021), with significant efforts underway stemming –importantly– from Africa itself (Nekoto et al., 2020).

To bring attention of the research community to the challenges of translating African Languages, this year we focus on the multilingual evaluation of 24 African Languages together with French and English. We base our evaluation on the benchmark provided by FLORES (Goyal et al., 2022), and its recent expansion to more African languages (NLLB Team et al., 2022).

In this second multilingual large-scale shared task, we evaluate the progress on massively multilingual translation for African Languages in a non-English-centric way. We propose 100 evaluation language pairs, based on regional clusters (South/South-East Africa, Horn of Africa and Central/East Africa, Nigeria and Gulf of Guinea, and Central Africa; and pivot languages (English, French). This year is characterized by two further innovations: first, a data track is included, in which participants share corpora to be used during the shared task; second, we perform human evaluation for a subset of the language pairs.

In the remainder of this paper, we describe the task setup, the participants, and the official results for the task. We also analyze the results to understand better the languages for which progress has been attained, and those where a gap in quality is still observed. Finally, we propose future directions for other tasks in the future.

Author names are sorted in alphabetical order.

[2] Progress was determined by comparing the average BLEU score between the best system in the task vs. the baseline.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 773 - 800
December 7-8, 2022 ©2022 Association for Computational Linguistics

2 Shared Task and Tracks

This year's task focuses on the 24 African languages listed in Table 1, along with French and English which are included as colonial linguae francae for evaluation purposes. These languages are all supported by the FLORES benchmark in its most recent expansion (NLLB Team et al., 2022).

Afrikaans	afr	Oromo	orm
Amharic	amh	Shona	sna
Chichewa	nya	Somali	som
Nigerian Fulfulde	fuv	Swahili	swh
Hausa	hau	Swati	ssw
Igbo	ibo	Tswana	tsn
Kamba	kam	Umbundu	umb
Kinyarwanda	kin	Wolof	wol
Lingala	lin	Xhosa	xho
Luganda	lug	Xitsonga	tso
Luo	luo	Yorùbá	yor
Northern Sotho	nso	Zulu	zul

Table 1: Focus African languages for this shared task. In addition to these, we also include French and English.

The human and automatic evaluation is based around 100 language directions, selected based on translator and annotator availability:

- **Languages of South and Southeast Africa**: xho-zul, zul-sna, sna-afr, afr-ssw, ssw-tsn, tsn-tso, tso-nso, nso-xho (8 directions).
- **Languages of the Horn of Africa**: swh-amh, amh-swh, luo-orm, som-amh, orm-som, swh-luo, amh-luo, luo-som (8 directions).
- **Languages of West Africa**: hau-ibo, ibo-yor, yor-fuv, fuv-hau, ibo-hau, yor-ibo, fuv-yor, hau-fuv, wol-hau, hau-wol, fuv-wol, wol-fuv (12 directions).
- **Languages of Central Africa**: kin-swh, lug-lin, nya-kin, swh-lug, lin-nya, lin-kin, kin-lug, nya-swh (8 directions).
- **Cross-regional pairs**: amh-zul, yor-swh, swh-yor, zul-amh, kin-hau, hau-kin, nya-som, som-nya, xho-lug, lug-xho, wol-swh, swh-wol (12 directions)
- **English pivots:** 22 languages translated into and from eng: afr, amh, nya, fuv, hau, ibo, kam, kin, lug, luo, nso, orm, sna, som, swh, ssw, tsn, umb, xho, tso, yor, zul (44 directions).
- **French pivots:** 4 languages translated into and from fra: kin, lin, swh, wol (8 directions).

3 Data Track

The data track focused on the contribution of novel corpora. Participants were welcomed to open-source and share monolingual, bilingual or multilingual datasets relevant to the training of MT models for this year's set of languages. There were seven submissions in this track:

LAVA[3] LAVA Corpus contains millions of parallel bilingual sentences, which are mined from Common Crawl. It covers five African languages.

MAFAND-MT (Adelani et al., 2022)[4] contains a few thousand high-quality and human translated parallel sentences for 21 African languages in the **news** domain. Each language has between 1,400 - 34,500 parallel sentences for training and/or evaluation. The languages covered are Amharic, Bambara, Ghomala, Ewe, Fon, Hausa, Igbo, Kinyarwanda, Luganda, Dholuo, Mossi, Chichewa, Nigerian-Pidgin, chiShona, Swahili, Setswana, Twi, Wolof, Yorùbá, isiXhosa, and isiZulu. These languages include 7 languages which were not present in the Shared task (and which are not reported therefore in Table 2).

KenTrans[5] This project produced a parallel corpus between Swahili and 2 other Kenya Languages: Dholuo and Luhya. The Luhya Language has several dialects. In the project 3 dialects were chosen as a start: Lumarachi, Logooli and Lubukusi. A total of 12,400 sentences were translated to Kiswahili from a sample of Dholuo and Luhya (1500 Dholuo-Kiswahili sentence pairs and 10,900 Luhya-Kiswahili sentence pairs). This corpus has an extension to speech in the version of Kencorpus (Wanjawa et al., 2022)[6].

Monolingual African languages from ParaCrawl[7] This release contains derived corpora built from language classified extracts of the ParaCrawl project. Monolingual data in this release comes from the Internet Archives and targeted crawls performed in the paracrawl project with document level language classification.

[3] https://drive.google.com/drive/folders/\179AkJ0P3fZMFS0rIyEBBDZ-WICs2wpWU

[4] https://github.com/masakhane-io/lafand-mt

[5] https://doi.org/10.7910/DVN/NOAT0W

[6] https://doi.org/10.7910/DVN/6N5V1K

[7] https://data.statmt.org/martin/

Dataset	African languages covered	No. of sentences	Participating Teams
LAVA	afr, kin, lug, nya, swa	3,225,801	Bytedance, GMU, ANVITA
MAFAND-MT	amh, hau, ibo, kin, lug, luo, nya, sna, swh, tsn, wol, xho, yor, zul	102,135	Bytedance, Tencent, DENTRA, ANVITA, Masakhane, GMU
KenTrans	luy, luo, swa	12,400	Bytedance, Tencent, DENTRA, ANVITA
ParaCrawl	afr, amh, fuv, hau, ibo, kam, lin, lug, luo, nso, nya, orm, sna, ssw, swh, tsn, tso, umb, wol, xho, yor, zul	22,349,179	Bytedance, Tencent, DENTRA, GMU
SA corpus	nso, tsn, xho, zul	160,035	Bytedance, Tencent, DENTRA, ANVITA
WebCrawlAfrican	afr, ling, ssw, amh, lug, tsn, nya, hau, orm, xho, ibo, tso, yor, swh, zul	695,000	Bytedance, Tencent, DENTRA, ANVITA

Table 2: Dataset submissions: covered Shared task languages, dataset sizes (i.e. number of parallel sentences in all translation directions), and participating teams that made use of the dataset in their submission.

SA Languages[8] The dataset was constructed using public available data mostly from South African Government websites.

WebCrawlAfrican (Vegi et al., 2022b) [9] Web Crawl African is a collection of African Multilingual parallel corpora comprising of 695,000 (approx) sentence pairs, covering 15 African languages plus English and 73 language pairs. African languages covered include Afrikaans, Lingala, Swati, Amharic, Luganda, Tswana, Chichewa, Hausa, Oroma, Xhosa, Igbo, Xitsonga, Yoruba, Swahili, Zulu. It covers variety of domains political, stories, religious and songs. Corpora have sentences covering both formal and informal writing styles.

This participation is summarised in Table 2 which includes language covered, dataset size and participating teams that currently used the dataset for their submission. All datasets where at least used by 3 teams in the evaluation. Note that the most used corpus was MAFAND-MT which was used by 6 different teams.

4 System Track

We provided a selection of training corpora, i.e. parallel sentences, to enable training of the MT systems. Submissions in the constrained translation track were only allowed to use data from the following sources:

- all corpora from the data track (see section 3);
- parallel corpora from OPUS (Tiedemann, 2012);[10]
- parallel corpora mined from Common Crawl using the LASER3 multilingual sentence encoder.

Participants who used other resources, had to submit to the unconstrained translation track.

Publicly available resources for African languages are very limited, for some of them less than fifty thousand sentences of bitexts are available. In addition to human translated sentences, several approaches were proposed to automatically mine parallel sentence from large collections of monolingual data. Unfortunately, recent approaches like ParaCrawl or CCMatrix (Schwenk et al., 2021) cover only few African languages. We extended the basic idea of mining based on a similarity measure in an multilingual embedding space (Artetxe and Schwenk, 2019) and developed sentence encoders for all African languages of this evaluation. We then performed bitext mining against 21.5 billion English sentences from Common Crawl, and 3.3 million sentences in French, respectively. These resources as well as the sentence encoders were made available to the participants of this evaluation. A detailed description of this mining approach can be found in Heffernan et al. (2022).

4.1 Participating Teams

CAIR ANVITA (Vegi et al., 2022a). The ANVITA-1.0 MT system is an English-centric multilingual transformer model for 24 African languages. The authors applied several heuristics to filter the data released for the shared task. They showed that that using larger Transformer model (24 encoder, 6 decoder) performs much better than smaller Transformer model (6 encoder, 6 decoder). Furthermore, they obtained some improvements by using an ensemble of last two epochs of Deep Transformer (24 encoder, 6 decoder).

[8] https://drive.google.com/drive/folders\/1jYwpxEdRxqXlB7BSmE6JxDar61U91xfI

[9] https://github.com/pavanpankaj/Web-Crawl-African

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

Cape Town (Elmadani et al., 2022). This system was focused on eight South and South-East African languages. The authors trained a multilingual NMT system (foreign to English and English to foreign, with some non English directions). The authors focused on exploring the best sub-word representation (BPE vs overlap-based BPE) and its effects on downstream performance for low-resource languages. Further, the authors explore creating synthetic data through back-translation and explore sampling techniques to balance the corpora.

GMU (Ibn Alam and Anastasopoulos, 2022). This system was based on on fine-tuning pretrained multilingual DeltaLM on 26 languages (625 translation directions). The fine-tuning was based on language- and language-family- (phylogeny) inspired adapter units (Faisal and Anastasopoulos, 2022) to improve its performance for African languages. The results show that a language-adapter-based fine-tuning significantly out-performs direct fine-tuning, but making use of family/sub-family adapters only helps in a few cases.

IIAI DenTra (Kamboj et al., 2022). This multilingual model combines a traditional translation loss with self-supervised tasks that can make use of unlabeled monolingual data. The resulting model performs denoising tasks (shuffling, masking) in conjunction with both translation and backtranslation. It then fine-tunes the model to in-domain data and covers 24 languages.

Masakhane (Abdulmumin et al., 2022). This model is based on M2M-100 which is fine-tuned on training data hat has been cleaned by an auxiliary language models. The pre-trained language models were fine-tuned on *positive* samples (clean data) vs. *negative* samples coming from automatically aligned data. The authors find significant improvements from using the filtered data.

Tencent Borderline (Jiao et al., 2022). The borderline model is a large transformer model (1.02B params.) which is augmented with data for zero-shot pairs through tagged back-translation and self-translation. In addition, it uses distributionally robust optimization (DRO) to alleviate the data imbalance. Finally it also uses family language information to group target languages and finetune separate models for each group.

Bytedance VolcTrans (Qian et al., 2022). This is an unconstrained system that uses different sources of parallel data (constrained data, NLLB, self-procured data) and monolingual data (e.g. VOA news, Wikipedia). This model is also trained on data cleaned through a rich set of heuristic rules to prevent punctuation mismatches, overly short/long sentences, among others; together with an approach based on minimum description length (MDL) that removes noisy sentences. The data is augmented with back-translation coming from pivot languages (Eng/Fra). The model is trained with target language tags added to both the encoder and decoder inputs. Finally, it includes post-processing rules for Yoruba accents.

SRPH-DAI (Cruz and Sutawika, 2022). This model is based on mT5, with additional adapters fine-tuned to each translation task, and then merged using adapter fusion to perform task-composition. The model is trained over data that is filtered using a set of heuristics. This model doesn't use other data-augmentation techniques (e.g. BT).

4.2 Automatic Evaluation

We follow last year's shared task in relying in spBLEU (Goyal et al., 2022) due to the well-known limitations of traditional BLEU. In particular we use a sentencepiece (Kudo and Richardson, 2018) tokenizer trained on all FLORES-200, in the hope of producing a *universal* tokenizer that can adequately handle all languages we are dealing with. Finally, to compute BLEU, we apply SPM tokenization to the system output and the reference, and then calculate BLEU at the sentence-piece level. For readability, in this paper we use BLEU and spBLEU interchangeably.

We additionally report chrF++ (Popović, 2017), another metric relying on character n-gram F-score (chrF) alongside word-level unigram and bigram F-score. This metric has been shown to correlate particularly well with human judgments for languages with rich morphology.

For all results we rely on statistical significance tests, using paired bootstrap resampling (Koehn, 2004) with 1000 samples. We first rank all the systems with spBLEU then we take the highest-scored system as a baseline and compare it with systems that are below its rank. If the p-value is greater than 0.05 we bundle those systems together. If for a system the p-value is less than 0.05 that means we have found a statistically significantly worse system and we make that system the new baseline system and continue to go on.

4.3 Human Evaluation

A fixed sample of the outputs of the primary submissions was grouped by the language pairs and split into tasks for evaluation by crowd human annotators. Each task comprised the source sentence and two translations of the source sentence that were to be scored from 0 to 100 indicating the general quality of the translations.

The annotation was conducted via crowdsourcing on the Toloka platform [11] where a crowd labeling project was set up for each language pair.

Guidelines for Evaluation The guidelines for the scale were roughly based on the theory of levels of translation equivalence by Komissarov (1990) and were simplified in order to facilitate their usage by annotators without linguistic background . The crowd annotators were asked to evaluate the translations on a 0..100 continuous rating scale (Graham et al., 2013) where 0 was considered a very bad and 100 — a very good translation (see Figure 1). The scale represents a combined approach that requires the annotators to give each translation one score assessing both accuracy and fluency at the same time and taking into account factors like grammaticality, naturalness, conveying the same meaning, having the same communicative goal, representing the same situation, having the same style and preserving as many shades of meaning of the source sentence as possible. The annotators were asked to evaluate the sentences as a whole and not word by word, pay attention to unmotivated additions of omissions, grammatical mistakes and untranslated parts, check if set expressions and metaphors were translated according to the expressions used in the target language and use their feel for the language in general.

Annotator Selection We ran the crowd annotation projects in Toloka and used the following criteria for annotator selection:

- Location: Africa according to the evaluator's IP address
- Both languages of the respective pair should have been among the list of languages known by the annotator as they had specified in their Profile
- Wherever possible, we showed the projects only to those annotators who had passed the respective language tests built in Toloka. The platform asks the annotators to verify their language skills via tests to get access to tasks with this prerequisite and it had such tests for English and French.

Potential annotators did not know about the requirements and did not see the projects if they did not meet the criteria. This approach prevented them from deliberate manipulation of their user settings to get access to the projects.

Quality Control The annotators whose profile matched our requirements, were shown the guidelines. The guidelines were translated into the target language of the respective pair as another way of testing the annotator's language skills.

Then the annotators were given an exam to test their ability to complete the task. The exam consisted of five tasks structurally identical to the main ones (see Figure 1). Thus each task consisted of the source sentence, two translations and a continuous scale for scoring. The exams were generated automatically from the list of source sentences paired with target sentences. The target sentences for the exam comprised:

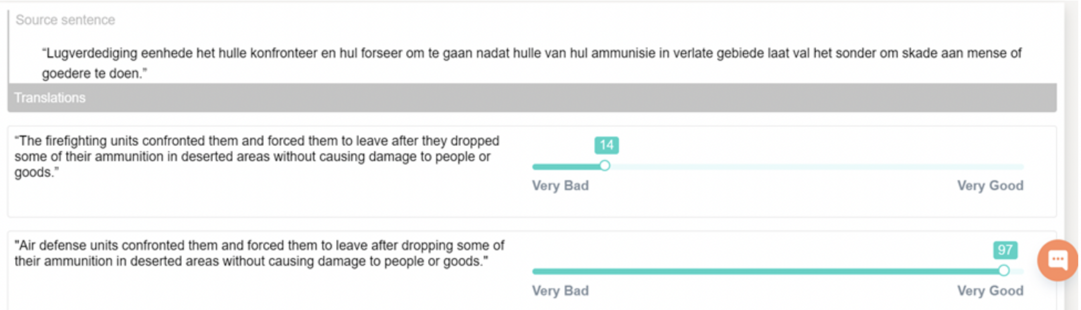
- Reference translations
- Reference translations with randomized word order (while still having correct capitalization when needed and ending with the respective punctuation mark)
- "Gibberish" sentences. Being potentially unknown to virtually any annotator, Sumerian was chosen as the source of "gibberish" control sentences [12] for eliminating cheaters. Furthermore, being unrelated to the languages in the dataset, it didn't confuse diligent annotators in terms of what pair of languages needs to be evaluated.

Each translation was assigned a golden score to be used as a control answer (Chida et al., 2022). Since the evaluation scale used is subjective and can show high variance, control intervals were introduced with the matching interval scores (score 0–33 = 0, 34–66 = 50, 66+ = 100). Randomized sentences as well as sentences in Sumerian were given golden scores 0 and reference translations were given a 100. Thus, if the annotator scored a bad sentence in a range of 0-33 it meant that they gave a correct answer and so on.

The task was accepted if both translations were given correct scores and rejected otherwise. The accuracy of the performer in an exam was calcu-

[11] https://toloka.ai

[12] https://etcsl.orinst.ox.ac.uk/cgi-bin/etcsl.cgi?text=c.1.8.1.1



Source sentence

"Lugverdediging eenhede het hulle konfronteer en hul forseer om te gaan nadat hulle van hul ammunisie in verlate gebiede laat val het sonder om skade aan mense of goedere te doen."

Translations

"The firefighting units confronted them and forced them to leave after they dropped some of their ammunition in deserted areas without causing damage to people or goods."

"Air defense units confronted them and forced them to leave after dropping some of their ammunition in deserted areas without causing damage to people or goods."

Figure 1: Screenshot of the interface with an annotated task comprising the source sentence and two translations randomly chosen from the outputs of the submitted models for the respective language pair (Afrikaans – English in the screenshot).

lated as the rate of accepted tasks among the five comprising the examination set. Exams where one of the languages (English or French) had been verified by the platform test required a 100% accuracy of completion (5/5 accepted tasks). For other languages the accuracy threshold was 80% to neutralize potential variance introduced by the automated creation of exams. These accuracy scores given in the exam were used as filters for the main pools of tasks, giving access to the main labeling only to the annotators who had passed the exam.

Another way of eliminating potential low-performers used both during the exam and the main set of tasks was banning the annotators for very fast responses. If an annotator submitted a page of five tasks within twenty seconds or less two times or more, they were banned from working on the project.

Score Normalization We convert the raw human scores to Z-scores:

$$z = \frac{x - \mu}{\sigma}$$

Note that we perform this operation at the annotator level; that means we compute the mean μ and standard deviation σ separately for each annotator and apply the transformation only on their scores.

Language Pairs Due to annotator availability, we only perform human evaluations in a smaller subset of our 100 language pairs. Hence the human score averages presented in all results only reflect averages for this subset, not all language pairs. The list of pairs is available in Table 15 in the Appendix.

5 Results

In this section we analyze both the human scores and automatic metrics for all the participants. We present the results and official ranking for the main task; the analysis of the performance from/to English, an to/from African languages; and the progress in performance w.r.t. to the previous year's task.

5.1 Main Results

The average results across the 100 evaluation language pairs are reported in Table 4. Note that we primarily only rank the *constrained* systems that were able to handle all language pairs.

System	# of pairs
Unconstrained	
Bytedance$_p$	52
Bytedance$_c$	48
Constrained	
Tencent$_p$	39
Tencent$_c$	26
GMU$_p$	17
DENTRA$_p$	9
GMU$_c$	7
Masakhane$_c$	1
SPRH-DAI$_c$	1
ANVITA, Masakhane$_p$ CapeTown, SPRH-DAI$_p$	0

Table 3: Number of evaluation language pairs (out of 100 total) that a system ranks (or ties for) best performance (spBLEU).

The best-performing constrained system on average is the Borderline (Tencent) system, followed by the GMU system with a little over 1 BLEU point difference between them. The only unconstrained submission is Bytedance's Volctrans, outperforming other systems by a significant margin of more than 4 BLEU points, but note however that it uses

Rank	System	Human	BLEU	spBLEU	chrF2
Systems handling all language pairs					
Constrained					
1	Tencent_p	0.39 ± 0.15	14.1 ± 1.1	17.6 ± 1.2	37.4 ± 1.3
	Tencent_c		14.0 ± 1.0	17.5 ± 1.2	37.2 ± 1.4
	GMU_c		13.3 ± 1.1	16.2 ± 1.1	35.4 ± 1.4
2	GMU_p	0.16 ± 0.28	13.3 ± 1.1	16.2 ± 1.1	35.4 ± 1.4
3	DENTRA_p	0.02 ± 0.51	10.4 ± 1.1	12.7 ± 1.2	30.5 ± 1.5
	SPRH-DAI_c		1.6 ± 1.2	2.0 ± 1.5	11.5 ± 0.4
4	SPRH-DAI_p	-1.4 ± 0.24	1.5 ± 1.6	1.8 ± 1.8	10.4 ± 0.5
Unconstrained					
U	ByteDance_p	0.53 ± 0.19	17.3 ± 1.2	21.9 ± 1.1	41.9 ± 1.2
U	ByteDance_c		17.3 ± 1.2	21.8 ± 1.1	41.7 ± 1.2
Systems handling partial language pairs					
P	ANVITA_p	0.24 ± 0.19	24.3 ± 1.1	26.3 ± 1.2	44.9 ± 1.2
P	ANVITA_c		23.8 ± 1.1	25.9 ± 1.2	44.5 ± 1.2
P	Capetown_p	0.09 ± 0.24	17.4 ± 1.0	21.3 ± 0.8	43.7 ± 0.7
P	Masakhane_p	0.05 ± 0.13	16.6 ± 1.0	19.0 ± 1.0	39.3 ± 1.1
P	Masakhane_c		11.9 ± 0.7	14.1 ± 0.7	35.0 ± 0.8

Table 4: Average results (presented here: mean ± standard error across all 100 evaluation language pairs), sorted by spBLEU. We only ranked constrained systems that handle all language pairs. We denote *unconstrained* submissions with U, and systems covering partial language pairs with P.

unconstrained data. All metrics, including human evaluation, result in a similar ranking among the 4 systems handling all language pairs.[13]

Table 3 lists the number of evaluation pairs for which each system ranks or ties for best performance. Among the constrained systems, the Tencent ones rank first for 39 language pairs, with the GMU one following with 17, and DENTRA with 9. The Masakhane and SPRH-DAI contrastive systems also rank first in one language pair each.

5.2 Performance by Language

Table 5 presents the best and worst performing language pairs. As shown, some language pairs end up with very good systems (e.g. Afrikaans, Swahili, Northern Sotho – at least when pairing with English). Wolof, Umbundu, and Kamba are the languages that most often show up as targets in the language pairs that systems struggle in, with BLEU scores below 10.

Results by for each individual language pair can be found in the Appendix Tables 16–115.

5.3 Performance by Groups

Table 6 presents the best performance across all systems, averaged over the different language groups we defined above. Note that having colonial languages (English and French) as the target perhaps unsurprisingly leads to generally high performance (average more than 33 BLEU). Languages from South Africa seem to be easier to translate (average ∼29 BLEU) while languages from West Africa, and especially Wolof, lead to worse performance as targets (average ∼15 BLEU).

Source	Target	spBLEU
eng	umb	4.1
eng	kam	5.9
fra	wol	8.3
swh	wol	8.5
hau	wol	8.6
afr	eng	60.1
swh	eng	49.0
eng	afr	46.0
nso	eng	43.5
eng	swh	42.0

Table 5: Top-5 worst (top) and best (bottom) language pairs (result from best system).

5.4 Progress from Last Year

To assess the progress made in African languages, we compare this year's best results with DeltaLM (Yang et al., 2021), the best system from last year's shared task. Note that this analysis only includes 72 of our 100 evaluation pairs, as some of the languages (e.g. Kinyarwanda or Swati) were

[13] For a fair comparison, Appendix Tables 13 and 14 present average scores for the partial language pairs that the Masakhane and CapeTown systems handle.

Group	as src	as trg	as both
South	29.4	22.8	19.9
Horn	21.9	20.0	16.5
West	20.3	15.1	14.4
Central	21.7	20.0	16.2
Colonial	22.3	33.7	N/A

Table 6: Average spBLEU of the best performing system summarized per language group.

added this year. In addition, note that last year's systems were scored using a different sentencepiece tokenizer than this year's ones.[14]

Across all 72 pairs, the average improvement is around 7.5 BLEU points. For around 93% of the 72 language pairs (67 pairs), there are improvements over last year's best system. We show the top-10 improved language pairs in Table 7. Most of the English-centric language pairs improve significantly, but african-to-african pairs benefit too, like the Swahili to Dholuo (swh-luo) pair which shows more than 12 BLEU points improvement. When indeed improving, the average improvement is more than 8 BLEU points (max: 33.5, min: 0.94).

Source	Target	2021	2022	Δ
nso	eng	9.9	43.5	33.5
eng	nso	9.5	30.5	21.1
yor	eng	7.1	25.1	18.0
hau	eng	22.2	40.0	17.8
orm	eng	10.6	28.3	17.7
eng	hau	16.4	31.5	15.1
som	eng	20.8	35.0	14.2
eng	luo	3.5	16.6	13.1
swh	luo	2.8	15.0	12.2
eng	som	8.7	20.8	12.1
Average (72 pairs)		15.1	22.6	7.5

Table 7: The top-10 language pairs with the largest improvement over last year's best result. Average refers to all 72 language pairs.

On the other hand, 5 language pairs do not improve from last year. Importantly, though, the largest drop is a mere 1.9 BLEU points for English to Afrikaans, and around 1 BLEU point reduction for the opposite direction as well as English to Kamba. For the few cases where we indeed observe a reduction, the average reduction is only 1 BLEU point (min: -0.46, max: -1.9).

Source	Target	2021	2022	Δ
eng	afr	47.9	46.0	-1.9
eng	kam	6.9	5.9	-1.0
afr	eng	61.0	60.1	-1.0
fra	lin	20.7	20.0	-0.7
eng	umb	4.6	4.1	-0.5

Table 8: The bottom-5 language pairs with the largest performance degradation over last year's best result.

5.5 X-eng and eng-X results

Rk	System	Human	BLEU	spBLEU	chrF2
Systems handling all language pairs					
Constrained					
1	Tencent$_p$	0.40	26.0	28.5	47.6
	Tencent$_c$		25.8	28.3	47.5
	GMU$_c$		25.9	28.0	46.6
2	GMU$_p$	0.22	25.8	28.0	47.0
3	ANVITA$_p$	0.24	24.3	26.3	44.9
	ANVITA$_c$		23.8	25.9	44.5
4	DENTRA$_p$	0.13	23.2	24.9	43.9
	SPRH-DA		2.5	3.2	14.0
5	SPRH-DA	-1.43	2.5	3.0	13.7
Unconstrained					
U	ByteDance$_p$	0.51	31.2	33.8	52.0
U	ByteDance$_c$		31.2	33.7	52.0
Systems handling partial language pairs					
P	Capetown$_p$	-0.04	25.3	27.4	47.8
P	Masakhane$_p$	-0.01	22.3	24.3	44.3
P	Masakhane$_c$		15.3	17.0	37.5

Table 9: Average results in all X-eng pairs, sorted by spBLEU. We only ranked constrained systems that handle all language pairs. We denote *unconstrained* submissions with U, and systems covering partial language pairs with P.

English-centric results (focusing on translating to and from English) are presented in Tables 10 and 9. The ranking of the system does not change depending on the direction. Note, however, that according to all metrics (including human evaluation) translating out of English and into the African languages is harder than the reverse direction for all systems.

5.6 Results on translation between African languages

Last, we present summary results on the average quality for translating between African languages,

[14] last year: spm tokenizer from FLORES-101; this year: spm tokenizer from FLORES-200.

Rk	System	Human	BLEU	spBLEU	chrF2
Systems handling all language pairs					
Constrained					
1	Tencent$_p$	0.33	14.1	18.8	39.5
	Tencent$_c$		13.9	18.5	39.1
2	DENTRA$_p$	0.22	12.8	16.7	37.5
3	GMU$_p$	0.02	12.0	15.2	35.3
	GMU$_c$		12.0	15.1	35.3
	SPRH-DAI$_c$		0.8	0.9	9.0
4	SPRH-DAI$_p$	-1.38	0.6	0.6	7.1
Unonstrained					
U	ByteDance$_p$	0.55	16.6	22.7	43.5
U	ByteDance$_c$		16.6	22.6	43.4
Systems handling partial language pairs					
P	Capetown$_p$	0.13	16.6	22.1	45.4
P	Masakhane$_p$	0.11	14.5	17.5	39.0
P	Masakhane$_c$		10.2	13.2	34.8

Table 10: Average results in all eng-X pairs, sorted by spBLEU. We only ranked constrained systems that handle all language pairs. We denote *unconstrained* submissions with U, and systems covering partial language pairs with P.

Rk	System	Human	BLEU	spBLEU	chrF2
Systems handling all language pairs					
Constrained					
1	Tencent$_p$	0.40	8.0	11.5	31.0
	Tencent$_c$		8.0	11.4	30.9
2	GMU$_p$	0.33	7.7	10.9	29.9
	GMU$_c$		7.7	10.8	29.9
3	DENTRA	-0.64	2.6	4.2	19.5
Unconstrained					
U	ByteDance$_p$	0.51	10.6	15.5	35.7
U	ByteDance$_c$		10.5	15.4	35.6
Systems handling partial language pairs					
P	Capetown$_p$	0.24	10.2	14.4	37.9

Table 11: Average results in all African-African pairs, sorted by spBLEU. We only ranked constrained systems that handle all language pairs. We denote *unconstrained* submissions with U, and systems covering partial language pairs with P.

shown in Table 11. It is worth noting that, although the automatic metrics score imply that the systems are much worse than in the English-centric directions (compare, for example, spBLEU scores of 22.7 on eng-X to 15.5 on african-to-arfican for the ByteDance system), the human evaluation scores are not too different. For instance, the ByteDance system receives an average human Z-score of 0.55 on eng-X and 0.51 on african-african; the Tencent system, with respective scores of 0.33 and 0.40, receives even higher Z-scores for african-to-african languages.

6 Conclusion and Future Work

In this paper, we presented the results of the second shared task on Large-Scale Machine Translation Evaluation. In this edition of the shared task, we evaluate the progress on massively multilingual translation for African Languages in a non-English-centric way. From our findings, we observe that data is still an important factor in translation performance, and that systems that used more data, either in an unconstrained way, or through data augmentation techniques made the most progress.The quality of the data is also important, as most participants developed a set of heuristics to clean it. We also observed the popularity of pre-trained translation models such as: DeltaLM, M2M-100 and mT5, as most systems used a version of these when developing their final model. We observe that there has been a large progress in the quality of translation in since the last iteration of this shared task: there is an improvement of about 7.5 BLEU points across 72 language pairs, and the average BLEU scores went from 15.09 to 22.60. We observe that it is usually harder to translate into than out of African Languages, and it is particularly difficult to translate into West African Languages like Wolof. We also observed that there has been significant progress translating into and out of Northern Sotho, Hausa, and Somali.

Acknowledgements

We thank Kevin Heffernan for developing LASER3 encoders, Onur Çelebi for work on mined bitext and Kenneth Heafield for work on mined bitext and open sourcing. We are also grateful to Toloka for running the human evaluation, as well as to the annotators who contributed with their expertise. We thank Dmitry Ustalov and Mariya Shmatova for sharing their expertise in human evaluation of machine translation. We thank Microsoft Research for contributing Azure compute credits for some participating teams. We are thankful to GMU's ORC computing cluster, where the automatic system test set generation and evaluation was hosted and ran. Mahfuz Alam, Fahim Faisal, and Antonios Anastasopoulos are generously supported by NSF awards FAI-2040926, IIS-2125466, and a Sponsored Research Award from Meta.

References

Idris Abdulmumin, Michael Coenraad Beukman, Jesujoba Alabi, Chris Chinenye Emezue, Everlyn Asiko Chimoto, Tosin Adewumi, Shamsuddeen Hassan Muhammad, Mofetoluwa Oluwaseun Adeyemi, Oreen Yousuf, Sahib Singh, and Tajuddeen Gwadabe. 2022. Separating grains from the chaff: Using data filtering to improve multilingual translation for low-resourced african languages. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Hiroki Chida, Yohei Murakami, and Mondheera Pituxcoosuvarn. 2022. Quality control for crowdsourced bilingual dictionary in low-resource languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6590–6596, Marseille, France. European Language Resources Association.

Jan Christian Blaise Cruz and Lintang Sutawika. 2022. Samsung research philippines - datasaur ai's submission for the wmt22 large scale multilingual translation task. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

Khalid N. Elmadani, Francois Meyer, and Jan Buys. 2022. University of cape town's WMT22 system: Multilingual machine translation for southern african languages. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. *JMLR*.

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

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

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *EMNLP*.

Md Mahfuz Ibn Alam and Antonios Anastasopoulos. 2022. Language adapters for large-scale mt: The GMU system for the wmt 2022 large-scale machine translation evaluation for african languages shared task. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

Wenxiang Jiao, Zhaopeng Tu, Jiarui Li, Wenxuan Wang, Jen tse Huang, and Shuming Shi. 2022. Tencent's multilingual machine translation system for wmt22 large-scale african languages. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

Ratnesh Joshi, Arindam Chatterjee, and Asif Ekbal. 2021. Towards explainable dialogue system: Explaining intent classification using saliency techniques. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 120–127, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Samta Kamboj, Sunil Kumar Sahu, and Neha Sengupta. 2022. DENTRA: Denoising and translation pre-training for multilingual machine translation. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Vilen N Komissarov. 1990. Theory of translation (linguistic aspects). *Moscow: Vysshaya shkola*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. DeltaLM: encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. `https://arxiv.org/abs/2106.13736`.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint 2207.04672*.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Z. Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan Van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020. Masakhane - machine translation for africa. *CoRR*, abs/2003.11529.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Xian Qian, Kai Hu, Jiaqiang Wang, Yifeng Liu, Xingyuan Pan, Jun Cao, and Mingxuan Wang. 2022. The volctrans system for wmt22 multilingual machine translation task. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. 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 6490–6500, Online. Association for Computational Linguistics.

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

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Prasanna Kumar K R, and Chitra Viswanathan. 2022a. ANVITA multilingual neural machine translation system for african languages. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Abhinav Mishra, Prashant Banjare, Prasanna Kumar K R, and Chitra Viswanathan. 2022b. Webcrawl african: A multilingual parallel corpora for african languages. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

Barack Wamkaya Wanjawa, Lilian D. A. Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, and Lawrence Muchemi. 2022. Kencorpus: A kenyan language corpus of swahili, dholuo and luhya for natural language processing tasks. *ArXiv*, abs/2208.12081.

Guillaume Wenzek, Vishrav Chaudhary, Angela Fan, Sahir Gomez, Naman Goyal, Somya Jain, Douwe Kiela, Tristan Thrush, and Francisco Guzmán. 2021. Findings of the WMT 2021 shared task on large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 89–99, Online. Association for Computational Linguistics.

Jian Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan, Xia Song, and Furu Wei. 2021. Multilingual machine translation systems from microsoft for wmt21 shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 446–455, Online. Association for Computational Linguistics.

A Appendix: Improvement over Last Year

Table 12 compares the performance of last year's best system to this year's best system for the 72 language pairs that intersect between the two.

Lang Pair	**2021**	**2022**	**Δ**	**Lang Pair**	**2021**	**2022**	**Δ**
eng-afr	47.9	46.0	-1.9	fra-wol	5.8	8.3	2.5
eng-amh	23.4	31.1	7.7	lin-fra	19.7	26.6	6.9
eng-hau	16.4	31.5	15.1	swh-fra	28.0	39.9	11.8
eng-ibo	17.9	23.5	5.6	wol-fra	11.2	21.6	10.4
eng-kam	6.9	5.9	-1.0	xho-zul	17.5	21.0	3.5
eng-lug	10.6	15.4	4.8	zul-sna	15.5	17.1	1.6
eng-luo	3.5	16.6	13.1	sna-afr	18.6	21.6	3.0
eng-nso	9.5	30.5	21.1	nso-xho	6.6	17.9	11.3
eng-nya	17.6	20.3	2.7	swh-amh	17.5	24.4	6.9
eng-orm	9.2	14.6	5.4	amh-swh	19.2	26.6	7.4
eng-sna	20.4	21.3	0.9	luo-orm	5.8	9.3	3.5
eng-som	8.7	20.8	12.1	som-amh	11.0	18.9	7.9
eng-swh	32.8	42.0	9.2	orm-som	3.3	12.9	9.6
eng-umb	4.6	4.1	-0.5	swh-luo	2.8	15.0	12.2
eng-xho	20.8	23.0	2.2	amh-luo	2.0	12.0	10.0
eng-yor	3.9	12.3	8.4	luo-som	5.0	12.7	7.7
eng-zul	22.3	28.4	6.1	hau-ibo	10.6	17.9	7.3
afr-eng	61.0	60.1	-1.0	ibo-yor	5.8	9.9	4.1
amh-eng	30.8	39.5	8.7	ibo-hau	10.3	21.6	11.3
hau-eng	22.2	40.0	17.8	yor-ibo	5.7	13.5	7.8
ibo-eng	25.3	36.4	11.1	wol-hau	6.6	14.6	8.0
kam-eng	11.2	18.3	7.1	hau-wol	4.6	8.6	4.0
lug-eng	16.6	26.2	9.6	lug-lin	13.3	14.8	1.5
luo-eng	20.0	27.5	7.5	swh-lug	8.5	13.0	4.5
nso-eng	10.0	43.5	33.5	lin-nya	11.5	13.8	2.3
nya-eng	23.0	32.7	9.7	nya-swh	16.2	23.0	6.8
orm-eng	10.6	28.3	17.7	amh-zul	14.0	18.7	4.7
sna-eng	25.5	31.8	6.4	yor-swh	5.8	17.9	12.1
som-eng	20.8	35.0	14.2	swh-yor	6.2	11.0	4.7
swh-eng	37.8	49.0	11.2	zul-amh	14.7	21.2	6.4
umb-eng	9.4	11.9	2.5	nya-som	5.3	13.9	8.6
xho-eng	30.7	38.3	7.6	som-nya	10.7	14.4	3.7
yor-eng	7.1	25.1	18.0	xho-lug	8.6	12.1	3.5
zul-eng	31.2	41.4	10.2	lug-xho	9.1	13.0	3.9
fra-lin	20.7	20.0	-0.7	wol-swh	8.5	16.5	8.0
fra-swh	24.7	30.8	6.1	swh-wol	5.6	8.5	2.9
				Average (72 pairs)	**15.09**	**22.60**	**7.51**

Table 12: Results on African languages on the FLORES-200 test set from last year to this year.

B Appendix: Comparisons for Masakhane and Capetown Models

Two submitted systems (Masakhane and Capetown) only handled some of our 100 evaluation language pairs. For a fair comparison, Tables 13 and 14 present average scores for the language pairs these systems handle. Overall, systems rankings do not change.

Rank	System	Human	BLEU	spBLEU	chrF2
Systems handling all language pairs					
U	ByteDance_pr*	0.54	25.1	28.9	48.2
U	ByteDance_contr*		25.1	28.8	48.1
1	Tencent_pr	0.37	21.3	24.5	44.6
	Tencent_contr		21.0	24.2	44.2
2	DENTRA_pr	0.23	19.1	21.8	42.2
	GMU_contr		19.5	21.5	41.0
3	GMU_pr	0.07	19.4	21.4	40.8
4	Masakhane_pr	0.05	16.6	19.0	39.3
	Masakhane_contr		11.9	14.1	35.0
Systems handling partial language pairs					
P	ANVITA_pr	0.23	26.9	29.0	48.2
P	ANVITA_contr		26.1	28.2	47.6
P	Capetown_pr	0.00	19.3	23.3	45.3
P	SPRH-DAI_contr		1.6	2.0	11.7
P	SPRH-DAI_pr	-1.42	1.4	1.8	10.6

Table 13: Average results in all pairs where Masakhane system participated, sorted by spBLEU. We only ranked constrained systems that handle all language pairs. We denote *unconstrained* submissions with U, and systems covering partial language pairs with P.

Rank	System	Human	BLEU	spBLEU	chrF2
Systems handling all language pairs					
U	ByteDance_contr*		24.2	29.5	50.0
U	ByteDance_pr*	0.41	24.1	29.5	50.1
1	Tencent_pr	0.41	21.9	27.0	48.6
	Tencent_contr		21.7	26.8	48.4
2	GMU_pr	0.21	20.7	24.8	46.0
	GMU_contr		20.7	24.7	46.0
3	Capetown_pr	0.09	17.4	21.3	43.7
4	DENTRA_pr	0.12	17.5	21.1	41.5
Systems handling partial language pairs					
P	ANVITA_pr	0.36	32.0	34.3	52.7
P	ANVITA_contr		31.6	33.9	52.4
P	Masakhane_pr	0.05	19.2	23.1	44.9
P	Masakhane_contr		11.1	14.0	35.8
P	SPRH-DAI_pr	-1.45	2.2	2.6	12.8
P	SPRH-DAI_contr		2.1	2.5	13.0

Table 14: Average results in all pairs where CapeTown system participated, sorted by spBLEU. We only ranked constrained systems that handle all language pairs. We denote *unconstrained* submissions with U, and systems covering partial language pairs with P.

C Appendix: Human Evaluation Guidelines

C.1 About

In the interface you will see a source sentence and two translations. Your task is to evaluate the quality of translations on a 0-100 scale where 0 is ridiculously bad and 100 is a perfect translation.

By doing this you will help us a lot to improve the quality of machine translation for African languages in all their glory and diversity.

C.2 How To Evaluate

The evaluation slider can go from 0 to 100. While choosing the most appropriate score, please consider the following features of a good translation starting from the most important:

1. Acceptable translation is grammatically correct, looks natural and makes sense to the reader.
2. Acceptable translation also has the same general meaning and communicative goal (what did it want to say?) as the source sentence
3. OK translation must be completely fluent and natural in addition to the above
4. Good translation keeps the style of the source sentence (e.g. formal or informal, colloquial style) in addition to being fluent and conveying the same meaning
5. Great translation keeps as much meaning and nuances as the source sentence in addition to being perfectly fluent.
6. Amazing translation also chooses the same means to describe the situation as the source sentence wherever the destination language has the same ways of doing it: the same metaphors, set expressions and such.

While performing the task, please do not use any automatic translation as the goal is to evaluate it with the help of human experts. You can use a dictionary if you found a word that you don't know.

C.2.1 Tips

- Evaluate the phrases as a whole and not word by word
- Check if any meaning was lost or unnecessarily added
- Check if there are any grammatical mistakes
- Check if anything remained untranslated
- Check if set expressions and metaphors were translated simply word by word (bad) or as a whole and according to the expressions used in the destination language (good)
- Use your own feeling as the speaker: do you consider the translation good, natural, clear and easily understandable

Language Pairs used in Evaluation			
eng-afr	eng-xho	som-eng	swh-fra
eng-amh	eng-yor	ssw-eng	xho-zul
eng-hau	eng-zul	swh-eng	zul-sna
eng-ibo	afr-eng	tsn-eng	afr-ssw
eng-lug	amh-eng	tso-eng	ssw-tsn
eng-nya	hau-eng	xho-eng	tsn-tso
eng-orm	ibo-eng	yor-eng	hau-ibo
eng-kin	lug-eng	zul-eng	ibo-yor
eng-sna	nya-eng	fra-lin	ibo-hau
eng-ssw	orm-eng	fra-swh	yor-ibo
eng-swh	kin-eng	fra-wol	swh-lug
eng-tsn	sna-eng	lin-fra	

Table 15: List of languages used for human evalulation

D Appendix: Individual Language Pair Results

System	Rank	spBLEU	BLEU	chrF2
Tencent_pr	1-3	46.0	40.2	68.0
ByteDance_pr	1-3	45.9	39.6	67.9
ByteDance_contr	1-3	45.7	39.3	67.8
Tencent_contr	4-6	45.5	40.0	67.6
DENTRA_pr	4-6	45.2	39.7	67.8
GMU_contr	4-6	45.2	39.8	67.5
GMU_pr	7	44.9	39.6	67.3
CapeTown_pr	8	40.4	35.9	64.3
SPRH-DAI_pr	9	4.3	4.0	21.8
SPRH-DAI_contr	10	3.0	2.6	19.6

Table 16: Results in eng-afr, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	31.1	12.9	42.4
ByteDance_pr	1-2	30.9	12.5	42.3
Tencent_pr	3	23.5	8.2	37.6
GMU_contr	4-5	22.0	7.6	35.5
Tencent_contr	4-5	21.9	7.7	36.5
GMU_pr	6	21.4	7.3	35.0
DENTRA_pr	7	15.2	5.2	29.4
SPRH-DAI_contr	8	0.2	0.3	0.5
SPRH-DAI_pr	9	0.0	0.1	2.9

Table 17: Results in eng-amh, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	5.4	3.8	24.7
ByteDance_contr	1-2	5.3	3.8	24.5
SPRH-DAI_contr	3-4	1.2	0.9	11.9
DENTRA_pr	3-4	1.1	0.7	14.0
Tencent_pr	5-6	0.9	0.5	14.8
Tencent_contr	5-6	0.8	0.4	14.7
GMU_pr	7-8	0.5	0.3	14.4
GMU_contr	7-8	0.5	0.3	14.0
SPRH-DAI_pr	9	0.1	0.1	4.8

Table 18: Results in eng-fuv, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	31.4	28.3	54.9
ByteDance_contr	2	30.9	27.9	54.8
Tencent_contr	3-4	28.7	25.4	53.9
Tencent_pr	3-4	28.7	25.5	53.9
DENTRA_pr	5	25.0	22.9	51.4
Masakhane_pr	6	19.7	17.7	45.8
Masakhane_contr	7	12.2	10.7	38.4
GMU_pr	8	5.2	14.5	30.0
GMU_contr	9	4.5	13.3	27.7
SPRH-DAI_contr	10	1.0	0.5	10.8
SPRH-DAI_pr	11	0.5	0.3	9.7

Table 19: Results in eng-hau, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
Tencent_pr	1	23.6	20.1	45.9
ByteDance_pr	2-4	23.2	19.7	45.6
ByteDance_contr	2-4	23.0	19.6	45.5
Tencent_contr	2-4	23.0	19.6	45.3
GMU_pr	5-6	19.6	17.3	42.4
GMU_contr	5-6	19.6	17.3	42.7
DENTRA_pr	7	18.0	16.4	42.0
Masakhane_pr	8	17.7	15.2	40.4
Masakhane_contr	9	14.7	11.9	36.5
SPRH-DAI_contr	10	0.8	0.6	10.4
SPRH-DAI_pr	11	0.1	0.2	5.0

Table 20: Results in eng-ibo, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	5.9	4.7	26.2
ByteDance_contr	2	5.3	4.1	25.4
GMU_pr	3-4	4.0	3.0	21.6
GMU_contr	3-4	4.0	2.9	21.6
DENTRA_pr	5-6	3.0	2.5	20.8
Tencent_pr	5-6	2.8	2.0	20.9
Tencent_contr	7	2.5	1.8	20.2
SPRH-DAI_contr	8	0.8	0.6	9.3
SPRH-DAI_pr	9	0.1	0.1	3.0

Table 21: Results in eng-kam, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	15.4	9.8	45.5
ByteDance_contr	2	15.2	9.7	45.4
DENTRA_pr	3-5	7.8	5.2	35.2
Tencent_pr	3-5	7.8	5.9	36.5
GMU_contr	3-5	7.5	5.8	34.8
GMU_pr	6-7	7.0	5.6	33.6
Tencent_contr	6-7	7.0	5.5	35.2
Masakhane_contr	8	6.2	4.5	33.1
Masakhane_pr	9	5.4	4.6	31.0
SPRH-DAI_contr	10	0.9	1.1	10.0
SPRH-DAI_pr	11	0.3	0.2	4.7

Table 22: Results in eng-lug, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	16.6	12.6	42.6
ByteDance_contr	2	16.4	12.4	42.4
GMU_pr	3-4	10.4	8.1	33.8
GMU_contr	3-4	10.4	8.0	33.9
DENTRA_pr	5	7.7	6.1	29.7
Tencent_pr	6	6.5	4.9	28.3
Tencent_contr	7	5.7	4.4	27.2
SPRH-DAI_contr	8	1.4	1.0	11.9
SPRH-DAI_pr	9	0.6	0.5	6.3

Table 23: Results in eng-luo, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	30.5	27.9	54.6
ByteDance_pr	1-2	30.4	27.7	54.6
Tencent_pr	3	28.5	25.4	53.8
Tencent_contr	4	28.0	24.9	53.3
DENTRA_pr	5	26.3	24.7	51.5
GMU_contr	6	24.8	23.5	49.9
GMU_pr	7-8	24.4	23.0	49.1
CapeTown_pr	7-8	24.1	22.7	50.0
SPRH-DAI_contr	9	0.9	0.4	9.7
SPRH-DAI_pr	10	0.4	0.3	6.0

Table 24: Results in eng-nso, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-3	20.3	15.4	50.2
ByteDance_contr	1-3	20.2	15.4	50.2
Tencent_contr	1-3	20.0	15.6	51.4
Tencent_pr	4	19.8	15.3	51.3
GMU_pr	5-6	17.2	13.4	48.4
GMU_contr	5-6	17.2	13.3	48.5
DENTRA_pr	7	16.3	13.3	48.0
SPRH-DAI_contr	8	1.4	1.4	13.2
SPRH-DAI_pr	9	0.8	1.1	12.2

Table 25: Results in eng-nya, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	14.6	6.7	45.1
ByteDance_pr	1-2	14.6	6.8	45.0
DENTRA_pr	3	4.4	2.1	28.3
Tencent_pr	4	2.8	1.4	25.2
GMU_pr	5-7	2.4	1.5	21.4
GMU_contr	5-7	2.4	1.4	21.6
Tencent_contr	5-7	2.4	1.2	23.4
SPRH-DAI_contr	8	0.1	0.1	5.8
SPRH-DAI_pr	9	0.0	0.0	2.9

Table 26: Results in eng-orm, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	29.6	23.3	55.5
ByteDance_contr	2	29.4	23.2	55.2
Tencent_pr	3	22.6	18.1	49.4
Tencent_contr	4	21.9	17.8	48.7
DENTRA_pr	5	18.5	14.4	45.8
GMU_contr	6	16.4	13.2	41.7
GMU_pr	7	16.2	12.8	41.2
SPRH-DAI_contr	8	0.4	0.4	9.6
SPRH-DAI_pr	9	0.4	0.3	6.4

Table 27: Results in eng-kin, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1	21.3	13.3	47.8
ByteDance_pr	2-3	21.1	13.1	47.8
Tencent_pr	2-3	20.8	12.9	49.3
Tencent_contr	4	20.5	12.7	49.1
DENTRA_pr	5	18.8	11.9	47.4
CapeTown_pr	6	17.6	10.3	46.4
GMU_contr	7-8	16.8	10.6	46.1
GMU_pr	7-8	16.7	10.6	46.1
SPRH-DAI_contr	9	0.8	1.0	10.8
SPRH-DAI_pr	10	0.8	1.1	13.9

Table 28: Results in eng-sna, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	20.8	15.0	48.2
ByteDance_pr	1-2	20.8	14.9	48.2
Tencent_pr	3	17.8	12.2	47.1
GMU_contr	4-6	17.5	11.9	45.8
Tencent_contr	4-6	17.5	11.9	46.9
GMU_pr	4-6	17.3	11.9	45.7
DENTRA_pr	7	15.1	10.4	43.5
SPRH-DAI_contr	8	0.5	0.4	8.8
SPRH-DAI_pr	9	0.1	0.3	8.0

Table 29: Results in eng-som, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	22.2	11.1	52.2
ByteDance_pr	1-2	22.0	11.3	52.6
Tencent_pr	3	19.2	9.6	49.4
Tencent_contr	4	18.8	9.5	48.9
DENTRA_pr	5	16.4	8.3	46.0
CapeTown_pr	6	15.5	7.6	44.9
GMU_contr	7	15.1	7.2	45.4
GMU_pr	8	14.4	6.8	44.4
SPRH-DAI_contr	9-10	0.8	1.1	10.8
SPRH-DAI_pr	9-10	0.8	0.7	10.9

Table 30: Results in eng-ssw, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	42.0	37.4	63.8
ByteDance_pr	1-2	42.0	37.2	63.7
Tencent_pr	3	39.2	34.8	62.8
Tencent_contr	4	38.8	34.4	62.5
DENTRA_pr	5	37.2	33.3	61.6
GMU_contr	6	36.5	32.7	61.2
GMU_pr	7	36.2	32.6	60.8
Masakhane_pr	8	35.1	31.5	60.2
Masakhane_contr	9	27.8	24.3	55.0
SPRH-DAI_contr	10	1.6	1.1	15.2
SPRH-DAI_pr	11	1.4	1.3	17.0

Table 31: Results in eng-swh, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	28.5	25.9	53.2
ByteDance_pr	1-2	28.4	25.9	53.1
Tencent_pr	3-4	25.6	22.9	50.3
Tencent_contr	3-4	25.3	22.7	49.9
DENTRA_pr	5	21.2	20.1	46.7
GMU_contr	6	20.7	19.7	46.0
GMU_pr	7	20.1	19.1	45.3
CapeTown_pr	8	19.7	18.8	45.3
Masakhane_pr	9	19.0	17.8	44.2
Masakhane_contr	10	11.0	10.1	36.0
SPRH-DAI_contr	11	0.7	0.3	9.1
SPRH-DAI_pr	12	0.3	0.3	5.9

Table 32: Results in eng-tsn, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	28.0	23.3	53.1
ByteDance_pr	1-2	27.8	23.0	53.2
Tencent_contr	3-4	21.7	18.8	49.5
Tencent_pr	3-4	21.4	18.8	49.6
GMU_contr	5-6	20.5	17.4	47.6
GMU_pr	5-6	20.1	17.2	47.2
DENTRA_pr	7	19.2	16.9	45.8
CapeTown_pr	8	17.9	15.8	44.7
SPRH-DAI_contr	9	1.1	0.7	10.5
SPRH-DAI_pr	10	0.4	0.3	4.2

Table 33: Results in eng-tso, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	4.1	2.0	31.3
ByteDance_contr	2	4.0	2.0	30.3
GMU_pr	3-4	2.2	0.9	23.3
GMU_contr	3-4	2.1	0.8	22.8
DENTRA_pr	5	2.0	1.2	22.9
Tencent_pr	6	1.8	0.9	22.6
Tencent_contr	7	1.6	0.9	21.8
SPRH-DAI_contr	8	0.7	0.6	9.8
SPRH-DAI_pr	9	0.2	0.3	3.7

Table 34: Results in eng-umb, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	23.0	12.1	51.2
ByteDance_pr	1-2	22.9	12.1	51.0
Tencent_contr	3	22.3	11.7	52.1
Tencent_pr	4	21.9	11.4	52.0
DENTRA_pr	5	20.2	10.2	50.1
CapeTown_pr	6	18.6	9.4	48.7
GMU_pr	7	3.5	1.4	20.2
GMU_contr	8	2.7	1.0	17.5
SPRH-DAI_pr	9	0.9	0.6	13.5
SPRH-DAI_contr	10	0.6	0.6	13.4

Table 35: Results in eng-xho, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	12.3	5.6	28.2
ByteDance_pr	1-2	12.3	5.5	28.2
Masakhane_contr	3	7.7	4.2	23.5
Tencent_pr	4	6.5	3.4	22.7
Tencent_contr	5	5.7	3.0	21.8
Masakhane_pr	6-8	5.0	3.2	21.5
GMU_pr	6-8	4.9	3.2	21.9
GMU_contr	6-8	4.8	3.2	21.8
DENTRA_pr	9	4.4	3.1	21.8
SPRH-DAI_contr	10	0.3	0.3	7.5
SPRH-DAI_pr	11	0.0	0.1	5.1

Table 36: Results in eng-yor, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	28.4	16.4	54.9
ByteDance_contr	1-2	28.3	16.4	54.9
Tencent_pr	3	26.8	14.9	55.3
Tencent_contr	4	26.4	14.5	55.0
GMU_pr	5-7	24.0	13.5	53.5
DENTRA_pr	5-7	24.0	12.7	53.1
GMU_contr	5-7	23.9	13.2	53.6
CapeTown_pr	8	22.8	11.9	52.0
Masakhane_pr	9	20.9	11.0	50.6
Masakhane_contr	10	13.1	6.0	41.8
SPRH-DAI_pr	11	0.8	0.5	13.0
SPRH-DAI_contr	12	0.5	0.6	12.4

Table 37: Results in eng-zul, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-4	60.1	57.4	75.5
ByteDance_pr	1-4	60.0	57.4	75.6
GMU_contr	1-4	60.0	56.9	76.1
GMU_pr	1-4	59.9	57.0	76.1
ANVITA_contr	5	58.7	55.7	75.5
Tencent_pr	6	58.1	55.0	75.2
DENTRA_pr	7-8	57.4	54.6	74.5
Tencent_contr	7-8	57.4	54.3	74.8
CapeTown_pr	9	46.4	44.6	67.4
SPRH-DAI_pr	10	9.0	8.3	27.3
SPRH-DAI_contr	11	7.1	6.1	22.9

Table 38: Results in afr-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	39.5	36.3	61.6
ByteDance_contr	2	39.2	36.2	61.4
GMU_contr	3	32.2	30.7	55.7
GMU_pr	4	31.6	30.1	55.3
Tencent_contr	5	29.0	27.6	53.6
Tencent_pr	6	28.6	26.7	53.2
ANVITA_contr	7	25.2	24.1	49.4
DENTRA_pr	8	23.3	22.5	48.4
SPRH-DAI_contr	9	1.0	0.9	12.4
SPRH-DAI_pr	10	0.8	0.7	12.2

Table 39: Results in amh-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	13.8	11.4	33.4
ByteDance_contr	1-2	13.7	11.3	33.2
GMU_pr	3-6	8.5	6.7	24.3
GMU_contr	3-6	8.5	6.9	25.0
Tencent_pr	3-6	8.5	6.5	25.2
DENTRA_pr	3-6	8.2	6.6	24.2
Tencent_contr	7-8	8.1	6.2	25.0
ANVITA_contr	7-8	7.9	6.1	23.4
SPRH-DAI_contr	9	2.0	1.4	12.1
SPRH-DAI_pr	10	1.8	1.3	11.7

Table 40: Results in fuv-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	40.0	37.7	58.8
ByteDance_contr	1-2	39.8	37.6	58.7
Tencent_pr	3	33.5	30.6	54.9
Tencent_contr	4-5	33.2	30.3	54.7
GMU_contr	4-5	32.4	29.6	52.6
GMU_pr	6-7	31.8	29.1	52.1
ANVITA_contr	6-7	31.3	28.8	52.3
DENTRA_pr	8	30.4	28.2	51.5
Masakhane_pr	9	25.1	22.7	46.5
Masakhane_contr	10	17.3	15.6	39.4
SPRH-DAI_pr	11-12	3.7	2.7	15.8
SPRH-DAI_contr	11-12	3.6	2.5	15.7

Table 41: Results in hau-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	36.4	33.6	56.7
ByteDance_pr	1-2	36.4	33.5	56.8
GMU_pr	3-4	30.8	28.2	51.6
GMU_contr	3-4	30.8	28.3	51.8
Tencent_contr	5	29.1	26.8	51.7
Tencent_pr	6	28.6	26.3	51.2
ANVITA_contr	7	25.8	23.6	47.5
DENTRA_pr	8	25.2	22.6	47.1
Masakhane_pr	9	23.2	20.9	45.9
Masakhane_contr	10	16.8	15.0	39.6
SPRH-DAI_contr	11	3.0	2.1	13.7
SPRH-DAI_pr	12	2.6	1.9	13.2

Table 42: Results in ibo-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	18.3	15.7	35.9
ByteDance_pr	1-2	18.3	15.7	36.6
GMU_pr	3-4	13.2	10.8	30.6
GMU_contr	3-4	13.2	10.9	30.5
ANVITA_contr	5-6	12.4	10.3	29.9
Tencent_pr	5-6	12.3	9.9	30.5
DENTRA_pr	7-8	11.9	9.8	29.2
Tencent_contr	7-8	11.7	9.4	30.3
SPRH-DAI_contr	9-10	2.9	2.1	13.6
SPRH-DAI_pr	9-10	2.8	2.1	13.2

Table 43: Results in kam-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	26.2	24.2	46.8
ByteDance_contr	1-2	26.1	24.2	46.6
Tencent_pr	3-4	21.8	19.7	41.9
Tencent_contr	3-4	21.6	19.7	41.8
GMU_pr	5	19.0	17.2	38.3
GMU_contr	6-7	18.7	16.8	37.8
ANVITA_contr	6-7	18.5	16.5	38.0
DENTRA_pr	8	18.1	16.4	37.7
Masakhane_pr	9	16.9	14.9	36.8
Masakhane_contr	10	13.9	12.2	35.0
SPRH-DAI_contr	11	2.6	2.0	14.1
SPRH-DAI_pr	12	2.4	1.8	13.0

Table 44: Results in lug-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	27.5	24.8	48.6
ByteDance_pr	1-2	27.4	24.7	48.6
Tencent_pr	3	22.0	19.3	43.0
Tencent_contr	4-6	21.5	18.8	42.8
GMU_contr	4-6	21.2	19.2	41.0
GMU_pr	4-6	21.0	19.1	40.7
DENTRA_pr	7-8	19.6	17.9	39.2
ANVITA_contr	7-8	19.5	17.6	39.3
SPRH-DAI_contr	9	2.4	1.8	12.7
SPRH-DAI_pr	10	2.2	1.8	12.0

Table 45: Results in luo-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	43.5	41.2	61.0
ByteDance_contr	2	43.1	40.7	60.7
Tencent_pr	3	39.6	37.1	58.5
Tencent_contr	4	39.2	36.7	58.2
GMU_pr	5	37.1	35.2	56.2
GMU_contr	6	36.7	34.6	55.8
ANVITA_contr	7	35.5	33.7	54.5
DENTRA_pr	8	33.8	32.2	53.8
CapeTown_pr	9	28.0	26.5	49.3
SPRH-DAI_contr	10	4.1	3.1	16.1
SPRH-DAI_pr	11	3.6	2.8	15.3

Table 46: Results in nso-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	32.7	28.9	53.0
ByteDance_pr	1-2	32.6	28.9	52.9
GMU_pr	3-4	29.0	25.8	49.6
GMU_contr	3-4	28.8	25.8	49.5
Tencent_pr	5-6	28.0	24.6	48.9
Tencent_contr	5-6	27.9	24.4	49.1
ANVITA_contr	7	26.2	23.1	47.1
DENTRA_pr	8	25.3	22.7	46.3
SPRH-DAI_contr	9	4.1	3.1	17.1
SPRH-DAI_pr	10	3.9	3.0	16.5

Table 47: Results in nya-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	28.3	26.3	50.9
ByteDance_contr	1-2	28.0	26.0	50.7
Tencent_contr	3-4	17.6	16.6	40.9
Tencent_pr	3-4	17.6	16.6	40.6
GMU_contr	5-6	15.3	14.6	36.8
GMU_pr	5-6	15.2	14.6	36.6
ANVITA_contr	7-8	12.0	11.2	32.9
DENTRA_pr	7-8	11.8	11.3	32.8
SPRH-DAI_contr	9	0.9	0.6	10.2
SPRH-DAI_pr	10	0.7	0.5	9.5

Table 48: Results in orm-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	38.7	35.8	58.5
ByteDance_pr	1-2	38.7	36.0	58.6
Tencent_pr	3-4	32.8	30.6	53.7
Tencent_contr	3-4	32.6	30.5	53.6
GMU_contr	5	30.2	28.4	51.3
GMU_pr	6	30.0	28.3	51.1
ANVITA_contr	7	28.9	27.4	50.1
DENTRA_pr	8	26.1	24.8	47.2
SPRH-DAI_contr	9	3.4	2.7	16.0
SPRH-DAI_pr	10	3.1	2.3	15.1

Table 49: Results in kin-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1	31.8	28.1	52.0
ByteDance_pr	2	31.4	27.8	51.8
GMU_pr	3-5	29.4	26.3	49.8
GMU_contr	3-5	29.3	26.1	49.5
Tencent_pr	3-5	29.1	25.6	49.6
Tencent_contr	6	28.5	24.9	49.4
ANVITA_contr	7	27.6	24.6	47.7
DENTRA_pr	8	26.0	23.2	46.9
CapeTown_pr	9	22.1	18.7	44.6
SPRH-DAI_contr	10-11	3.7	3.0	16.6
SPRH-DAI_pr	10-11	3.7	3.0	16.1

Table 50: Results in sna-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	35.0	32.9	55.1
ByteDance_contr	2	34.6	32.5	54.7
Tencent_pr	3-4	28.4	26.5	49.8
Tencent_contr	3-4	28.2	26.4	49.8
GMU_pr	5-6	27.8	26.4	48.1
GMU_contr	5-6	27.8	26.4	48.0
DENTRA_pr	7	23.4	21.9	44.6
ANVITA_contr	8	22.2	20.6	42.7
SPRH-DAI_contr	9	3.0	2.3	15.0
SPRH-DAI_pr	10	2.5	2.0	13.6

Table 51: Results in som-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1	36.8	34.3	56.5
ByteDance_pr	2	36.5	33.9	56.2
Tencent_pr	3	32.9	30.1	53.0
Tencent_contr	4	32.3	29.5	52.8
GMU_pr	5-6	29.2	27.1	49.8
GMU_contr	5-6	29.0	27.1	49.5
ANVITA_contr	7-8	27.5	25.5	47.5
DENTRA_pr	7-8	27.5	25.8	48.5
CapeTown_pr	9	23.5	21.5	45.2
SPRH-DAI_pr	10	3.5	2.6	14.9
SPRH-DAI_contr	11	3.3	2.6	15.0

Table 52: Results in ssw-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	49.0	47.4	67.7
ByteDance_contr	2	48.6	47.0	67.4
GMU_pr	3-6	42.8	41.0	62.8
Tencent_pr	3-6	42.8	41.1	62.9
GMU_contr	3-6	42.7	41.1	62.8
Tencent_contr	3-6	42.7	40.9	62.8
ANVITA_contr	7	41.7	40.4	62.2
DENTRA_pr	8	39.8	38.9	60.7
Masakhane_pr	9	36.4	35.2	58.4
Masakhane_contr	10	28.2	27.5	51.7
SPRH-DAI_contr	11-12	4.5	3.9	18.8
SPRH-DAI_pr	11-12	4.4	4.1	19.0

Table 53: Results in swh-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1	34.4	31.8	54.5
ByteDance_pr	2	34.2	31.6	54.4
Tencent_pr	3	30.3	27.5	51.7
Tencent_contr	4	29.8	26.9	51.5
GMU_pr	5	29.3	26.6	50.3
GMU_contr	6	28.2	25.6	48.9
ANVITA_contr	7	27.5	25.4	48.5
DENTRA_pr	8	26.2	23.9	47.6
Masakhane_pr	9	23.5	21.2	45.1
CapeTown_pr	10	22.1	19.8	44.2
Masakhane_contr	11	11.3	9.7	33.3
SPRH-DAI_contr	12	3.3	2.6	15.1
SPRH-DAI_pr	13	2.9	2.3	14.4

Table 54: Results in tsn-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	35.6	32.8	54.6
ByteDance_pr	1-2	35.6	32.8	54.7
Tencent_contr	3-4	31.9	29.6	52.1
Tencent_pr	3-4	31.9	29.7	52.1
GMU_pr	5-6	30.4	28.3	50.6
GMU_contr	5-6	30.2	28.1	50.2
ANVITA_contr	7-8	27.2	25.3	47.3
DENTRA_pr	7-8	27.2	25.6	47.7
CapeTown_pr	9	21.8	20.3	43.2
SPRH-DAI_contr	10	3.0	2.4	14.3
SPRH-DAI_pr	11	2.8	2.1	13.4

Table 55: Results in tso-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	11.9	9.7	31.0
ByteDance_pr	1-2	11.8	9.7	31.3
GMU_pr	3	9.9	8.0	28.0
GMU_contr	4-6	9.6	7.7	27.7
Tencent_contr	4-6	9.4	7.0	27.7
Tencent_pr	4-6	9.4	7.1	27.6
ANVITA_contr	7	8.1	6.2	26.4
DENTRA_pr	8	7.4	5.8	25.1
SPRH-DAI_contr	9-10	1.5	0.9	12.3
SPRH-DAI_pr	9-10	1.5	1.0	11.8

Table 56: Results in umb-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	38.3	35.4	57.7
ByteDance_contr	1-2	38.2	35.3	57.6
Tencent_pr	3-5	34.2	31.1	54.6
Tencent_contr	3-5	34.1	31.0	54.4
GMU_contr	3-5	33.9	31.3	54.3
GMU_pr	6	33.7	31.3	54.2
ANVITA_contr	7	32.4	29.8	52.9
DENTRA_pr	8	30.8	28.8	51.8
CapeTown_pr	9	26.7	24.3	49.2
SPRH-DAI_pr	10-11	4.1	3.3	17.5
SPRH-DAI_contr	10-11	4.0	3.2	17.5

Table 57: Results in xho-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	25.1	22.9	46.4
ByteDance_contr	1-2	24.9	22.6	46.0
Tencent_contr	3-4	20.2	17.4	41.3
Tencent_pr	3-4	20.2	17.5	41.5
GMU_contr	5-6	19.7	17.6	40.2
GMU_pr	5-6	19.6	17.6	40.1
ANVITA_contr	7	17.9	15.8	38.5
DENTRA_pr	8	16.4	14.5	37.2
Masakhane_pr	9	16.1	14.2	36.9
Masakhane_contr	10	10.9	8.8	31.2
SPRH-DAI_contr	11	2.5	1.8	13.3
SPRH-DAI_pr	12	2.1	1.5	12.1

Table 58: Results in yor-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	41.4	39.2	60.4
ByteDance_pr	1-2	41.3	39.0	60.4
GMU_pr	3-5	36.8	34.6	56.9
GMU_contr	3-5	36.8	34.4	56.9
Tencent_pr	3-5	36.6	34.0	56.6
Tencent_contr	6	36.1	33.3	56.4
ANVITA_contr	7	34.9	32.5	55.2
DENTRA_pr	8	32.7	31.1	53.5
Masakhane_pr	9	29.0	26.8	50.5
CapeTown_pr	10	28.5	26.7	50.7
Masakhane_contr	11	20.4	18.5	43.1
SPRH-DAI_contr	12-13	3.5	2.9	16.5
SPRH-DAI_pr	12-13	3.5	2.9	16.4

Table 59: Results in zul-eng, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	25.3	18.4	52.1
ByteDance_contr	2	25.1	18.3	51.9
Tencent_pr	3-4	19.3	14.1	47.0
Tencent_contr	3-4	19.0	14.4	46.6
DENTRA_pr	5-6	14.9	10.7	41.8
GMU_contr	5-6	14.4	10.3	41.3
GMU_pr	7	13.9	10.1	40.4

Table 60: Results in fra-kin, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	20.0	16.6	50.8
ByteDance_contr	2	19.7	16.2	50.7
DENTRA_pr	3	17.1	14.7	45.9
Tencent_pr	4-5	16.4	14.1	46.1
Tencent_contr	4-5	16.3	13.8	45.8
GMU_pr	6-7	10.1	7.5	37.4
GMU_contr	6-7	10.1	7.2	37.1

Table 61: Results in fra-lin, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	30.7	25.0	56.3
ByteDance_contr	2	30.2	24.5	56.0
Tencent_pr	3	29.2	24.1	55.2
Tencent_contr	4	28.8	23.8	54.9
GMU_pr	5	27.6	23.4	53.2
GMU_contr	6	27.1	22.8	52.8
DENTRA_pr	7	24.9	21.2	50.8

Table 62: Results in fra-swh, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
DENTRA_pr	1	8.3	5.6	28.2
ByteDance_pr	2-3	7.3	5.2	27.6
ByteDance_contr	2-3	7.1	5.0	27.5
Masakhane_contr	4	6.3	4.4	27.4
Tencent_pr	5	4.8	3.9	23.2
Tencent_contr	6	4.2	3.4	22.0
GMU_pr	7	3.0	2.0	15.9
GMU_contr	8	2.6	1.8	14.0
Masakhane_pr	9	2.2	1.5	19.5

Table 63: Results in fra-wol, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	33.7	29.1	54.1
ByteDance_pr	1-2	33.7	29.1	54.2
Tencent_pr	3	27.2	23.6	49.4
Tencent_contr	4	26.8	23.1	49.2
GMU_pr	5	26.4	23.0	48.0
GMU_contr	6	26.0	22.7	47.8
DENTRA_pr	7	22.2	18.5	43.4

Table 64: Results in kin-fra, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	26.6	22.7	46.9
ByteDance_pr	1-2	26.5	22.6	46.8
Tencent_contr	3-4	23.8	19.7	44.8
Tencent_pr	3-4	23.6	19.6	44.6
GMU_contr	5-6	22.9	19.1	43.1
GMU_pr	5-6	22.8	18.8	42.7
DENTRA_pr	7	20.7	16.9	41.7

Table 65: Results in lin-fra, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	39.9	35.5	60.0
ByteDance_contr	2	39.6	35.2	59.7
GMU_contr	3-4	35.0	31.0	56.0
GMU_pr	3-4	34.9	30.8	56.0
Tencent_pr	5-6	33.6	29.2	55.1
Tencent_contr	5-6	33.4	29.1	55.0
DENTRA_pr	7	31.1	26.8	53.0

Table 66: Results in swh-fra, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	21.6	17.8	41.7
ByteDance_contr	2	21.2	17.6	41.2
Tencent_contr	3-5	14.6	12.0	36.0
Tencent_pr	3-5	14.5	12.1	35.8
DENTRA_pr	3-5	14.4	11.2	35.1
GMU_contr	6-7	13.2	10.5	31.1
GMU_pr	6-7	13.0	10.3	30.6
Masakhane_pr	8	9.2	7.5	29.7
Masakhane_contr	9	8.3	7.0	30.2

Table 67: Results in wol-fra, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-3	21.0	10.3	48.1
ByteDance_contr	1-3	20.9	10.4	47.9
Tencent_pr	1-3	20.5	9.9	49.2
Tencent_contr	4-5	20.4	9.8	49.0
GMU_contr	4-5	20.1	10.0	49.1
GMU_pr	6	20.0	9.9	48.9
CapeTown_pr	7	18.0	8.5	47.4
DENTRA_pr	8	10.6	4.1	38.2

Table 68: Results in xho-zul, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-5	17.1	10.1	44.1
ByteDance_pr	1-5	17.1	10.0	44.3
Tencent_contr	1-5	17.0	9.9	45.7
Tencent_pr	1-5	17.0	10.0	45.8
GMU_pr	1-5	16.7	9.9	45.8
GMU_contr	6	16.6	9.9	45.8
CapeTown_pr	7	15.0	8.5	44.2
DENTRA_pr	8	4.4	2.1	25.2

Table 69: Results in zul-sna, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	21.6	17.3	45.6
ByteDance_pr	1-2	21.5	17.2	45.4
GMU_pr	3-4	20.1	17.0	43.8
GMU_contr	3-4	20.0	17.0	43.8
Tencent_contr	5-6	19.1	15.7	43.1
Tencent_pr	5-6	19.0	15.5	43.2
CapeTown_pr	7	15.1	12.0	40.1
DENTRA_pr	8	13.9	11.2	38.1

Table 70: Results in sna-afr, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	18.9	9.2	48.7
ByteDance_contr	1-2	18.7	8.9	48.4
Tencent_pr	3	17.3	7.8	47.9
Tencent_contr	4	16.6	7.9	46.5
GMU_contr	5	14.7	6.7	45.2
GMU_pr	6	14.3	6.5	44.7
CapeTown_pr	7	11.2	5.4	40.0
DENTRA_pr	8	8.3	4.3	36.1

Table 71: Results in afr-ssw, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	21.6	19.1	46.8
ByteDance_pr	1-2	21.6	19.2	46.9
Tencent_contr	3-4	19.1	17.1	44.1
Tencent_pr	3-4	19.0	16.8	44.3
GMU_contr	5	17.7	16.5	43.2
GMU_pr	6	17.1	15.9	42.7
CapeTown_pr	7	15.4	14.4	41.2
DENTRA_pr	8	3.5	1.8	24.8

Table 72: Results in ssw-tsn, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	20.5	16.5	46.5
ByteDance_pr	1-2	20.5	16.5	46.5
Tencent_pr	3	16.9	13.9	44.3
GMU_contr	4-5	16.2	13.5	44.0
Tencent_contr	4-5	16.2	13.6	44.1
GMU_pr	6-7	15.4	13.0	43.8
CapeTown_pr	6-7	15.1	13.2	41.9
DENTRA_pr	8	4.1	2.4	24.6

Table 73: Results in tsn-tso, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-3	21.0	18.3	45.7
ByteDance_pr	1-3	20.9	18.4	45.9
Tencent_contr	1-3	20.3	17.6	44.6
Tencent_pr	4-5	19.7	17.4	44.4
GMU_pr	4-5	19.2	17.9	44.5
GMU_contr	6	18.9	17.8	44.4
CapeTown_pr	7	12.0	13.1	38.7
DENTRA_pr	8	5.6	3.6	26.2

Table 74: Results in tso-nso, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-3	17.9	8.8	46.0
ByteDance_pr	1-3	17.8	8.5	46.1
Tencent_contr	1-3	17.5	8.4	46.8
Tencent_pr	4	16.8	8.5	46.1
GMU_contr	5-6	16.0	8.5	46.2
GMU_pr	5-6	15.9	8.7	46.3
CapeTown_pr	7	13.7	6.6	42.7
DENTRA_pr	8	3.0	1.4	24.3

Table 75: Results in nso-xho, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	24.4	9.1	35.9
ByteDance_contr	2	23.9	8.6	35.5
Tencent_pr	3-5	18.6	6.0	31.8
GMU_contr	3-5	18.5	5.9	32.0
GMU_pr	3-5	18.3	5.8	32.0
Tencent_contr	6	18.2	5.5	31.6
DENTRA_pr	7	3.1	1.3	10.1

Table 76: Results in swh-amh, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	26.6	21.8	52.8
ByteDance_pr	1-2	26.5	21.8	52.7
GMU_pr	3-5	21.9	18.6	49.5
GMU_contr	3-5	21.7	18.5	49.5
Tencent_contr	3-5	21.6	18.3	49.7
Tencent_pr	6	21.0	17.7	49.0
DENTRA_pr	7	11.7	10.0	38.8

Table 77: Results in amh-swh, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	9.3	3.9	37.2
ByteDance_pr	1-2	9.3	3.9	37.2
Tencent_pr	3	1.9	0.9	23.1
Tencent_contr	4	1.7	0.8	22.1
GMU_pr	5-7	1.3	0.7	18.5
GMU_contr	5-7	1.3	0.7	18.5
DENTRA_pr	5-7	1.2	0.8	16.3

Table 78: Results in luo-orm, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	18.9	6.7	29.9
ByteDance_contr	1-2	18.7	6.6	29.6
Tencent_pr	3-4	13.7	4.4	25.7
GMU_pr	3-4	13.3	4.1	25.7
GMU_contr	5-6	13.2	4.1	25.5
Tencent_contr	5-6	13.2	4.3	25.3
DENTRA_pr	7	0.6	0.4	1.8

Table 79: Results in som-amh, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	12.9	8.5	39.7
ByteDance_contr	1-2	12.8	8.6	39.5
Tencent_contr	3-4	8.5	5.5	35.3
Tencent_pr	3-4	8.5	5.5	35.4
GMU_pr	5-6	7.9	5.4	33.5
GMU_contr	5-6	7.8	5.4	33.2
DENTRA_pr	7	1.3	0.9	21.8

Table 80: Results in orm-som, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	15.0	11.3	41.4
ByteDance_contr	2	14.7	11.2	40.9
GMU_pr	3-4	8.8	6.6	31.9
GMU_contr	3-4	8.7	6.5	31.9
Tencent_pr	5-6	5.3	3.8	26.2
Tencent_contr	5-6	5.2	3.8	25.6
DENTRA_pr	7	3.9	2.5	21.7

Table 81: Results in swh-luo, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	12.0	8.8	38.4
ByteDance_pr	1-2	12.0	8.7	38.6
GMU_contr	3	6.4	5.0	29.5
GMU_pr	4	6.1	4.7	29.1
DENTRA_pr	5	3.1	2.8	24.7
Tencent_contr	6-7	2.8	2.0	21.2
Tencent_pr	6-7	2.8	2.1	21.4

Table 82: Results in amh-luo, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	12.7	8.5	39.0
ByteDance_pr	1-2	12.6	8.5	39.0
GMU_contr	3-5	9.1	6.3	34.7
Tencent_contr	3-5	9.0	6.2	35.8
GMU_pr	3-5	8.9	6.1	34.2
Tencent_pr	6	8.5	5.6	34.8
DENTRA_pr	7	4.1	2.8	20.4

Table 83: Results in luo-som, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
Tencent_contr	1-4	17.9	14.6	39.7
ByteDance_contr	1-4	17.8	14.9	39.8
Tencent_pr	1-4	17.8	14.7	39.4
ByteDance_pr	1-4	17.7	14.9	39.8
GMU_pr	5-6	15.7	13.5	37.5
GMU_contr	5-6	15.7	13.4	37.4
DENTRA_pr	7	4.2	2.7	19.2

Table 84: Results in hau-ibo, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	10.0	4.6	25.5
ByteDance_contr	2	9.7	4.5	25.6
Tencent_pr	3-4	5.2	2.9	21.4
Tencent_contr	3-4	5.1	2.8	20.8
GMU_pr	5	4.0	2.5	20.5
GMU_contr	6	3.9	2.5	20.3
DENTRA_pr	7	2.3	1.2	13.8

Table 85: Results in ibo-yor, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	4.0	2.7	22.3
ByteDance_pr	1-2	4.0	2.7	22.4
DENTRA_pr	3	1.9	0.9	13.7
Tencent_pr	4-7	0.5	0.2	13.9
GMU_pr	4-7	0.4	0.3	13.6
GMU_contr	4-7	0.4	0.3	13.5
Tencent_contr	4-7	0.3	0.2	14.0

Table 86: Results in yor-fuv, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	9.5	7.3	32.1
ByteDance_contr	1-2	9.4	7.2	31.9
Tencent_contr	3-4	3.4	2.8	22.4
Tencent_pr	3-4	3.4	2.9	23.1
GMU_pr	5-7	3.1	2.5	18.8
DENTRA_pr	5-7	3.1	1.7	19.5
GMU_contr	5-7	3.1	2.7	19.6

Table 87: Results in fuv-hau, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	21.6	18.8	46.7
ByteDance_pr	1-2	21.5	18.7	46.3
Tencent_pr	3-4	17.8	15.6	44.2
Tencent_contr	3-4	17.1	15.0	44.1
GMU_pr	5	17.0	14.9	42.6
GMU_contr	6	16.7	14.7	42.3
DENTRA_pr	7	3.8	2.2	20.0

Table 88: Results in ibo-hau, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1	13.5	10.9	35.0
ByteDance_pr	2	13.2	10.6	34.7
Tencent_contr	3	12.6	9.5	33.1
Tencent_pr	4	12.2	9.3	33.1
GMU_pr	5-6	11.5	9.3	32.2
GMU_contr	5-6	11.5	9.3	32.3
DENTRA_pr	7	2.4	1.0	13.1

Table 89: Results in yor-ibo, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	5.7	2.3	20.1
ByteDance_contr	1-2	5.6	2.2	19.8
DENTRA_pr	3	2.0	1.2	13.4
GMU_pr	4-5	1.0	0.6	9.2
Tencent_pr	4-5	0.9	0.4	10.2
Tencent_contr	6-7	0.8	0.4	9.4
GMU_contr	6-7	0.7	0.4	8.0

Table 90: Results in fuv-yor, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	5.0	3.5	23.8
ByteDance_contr	2	4.7	3.4	23.5
DENTRA_pr	3	3.0	1.5	21.1
GMU_pr	4	0.4	0.3	13.9
GMU_contr	5-7	0.4	0.3	13.6
Tencent_contr	5-7	0.4	0.2	14.0
Tencent_pr	5-7	0.4	0.1	14.0

Table 91: Results in hau-fuv, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	14.6	12.1	38.4
ByteDance_contr	1-2	14.4	11.9	38.1
Tencent_contr	3	9.1	7.6	34.1
Tencent_pr	4	8.5	7.2	32.8
GMU_pr	5	7.2	5.8	25.4
GMU_contr	6	6.9	5.7	26.4
DENTRA_pr	7	3.8	2.3	20.3

Table 92: Results in wol-hau, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	8.6	5.9	27.5
ByteDance_contr	2	8.2	5.6	27.1
Tencent_pr	3-4	3.8	3.1	20.0
GMU_pr	3-4	3.7	2.4	15.7
GMU_contr	5-6	3.5	2.3	14.9
Tencent_contr	5-6	3.4	2.7	18.8
DENTRA_pr	7	3.2	1.6	19.2

Table 93: Results in hau-wol, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	5.2	3.4	22.3
ByteDance_contr	2	5.0	3.3	21.9
DENTRA_pr	3	2.5	1.4	18.4
GMU_pr	4	1.3	0.9	10.7
GMU_contr	5	1.1	0.8	10.4
Tencent_pr	6-7	1.0	0.9	10.8
Tencent_contr	6-7	0.9	0.9	10.6

Table 94: Results in fuv-wol, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	4.4	3.0	23.0
ByteDance_contr	2	4.2	2.8	22.5
DENTRA_pr	3	2.5	1.3	19.3
GMU_pr	4-7	0.4	0.3	14.3
GMU_contr	4-7	0.4	0.3	13.9
Tencent_pr	4-7	0.4	0.3	14.4
Tencent_contr	4-7	0.3	0.2	14.6

Table 95: Results in wol-fuv, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	27.9	23.5	52.9
ByteDance_contr	2	27.5	23.1	52.7
Tencent_contr	3-4	24.2	20.8	50.7
Tencent_pr	3-4	24.2	20.8	50.4
GMU_contr	5-6	22.6	19.3	49.4
GMU_pr	5-6	22.4	19.3	49.0
DENTRA_pr	7	4.5	3.7	25.2

Table 96: Results in kin-swh, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	14.9	11.9	41.2
ByteDance_contr	1-2	14.8	11.8	41.0
Tencent_pr	3-4	9.1	7.9	32.7
Tencent_contr	3-4	9.0	7.8	32.9
GMU_contr	5-6	7.2	5.7	32.5
GMU_pr	5-6	7.0	5.5	32.0
DENTRA_pr	7	3.2	1.7	22.9

Table 97: Results in lug-lin, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	18.5	13.3	44.2
ByteDance_contr	1-2	18.3	13.2	44.1
Tencent_contr	3	15.8	11.9	41.9
Tencent_pr	4	15.0	10.9	41.5
GMU_contr	5	12.4	9.4	39.0
GMU_pr	6	11.9	9.0	38.5
DENTRA_pr	7	3.7	2.5	21.9

Table 98: Results in nya-kin, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	13.0	7.9	43.3
ByteDance_pr	1-2	13.0	8.0	43.6
Tencent_pr	3	8.4	5.9	37.1
Tencent_contr	4-5	7.5	5.6	35.9
GMU_pr	4-5	7.2	5.5	34.5
GMU_contr	6	6.5	5.0	33.6
DENTRA_pr	7	3.0	1.9	21.9

Table 99: Results in swh-lug, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	13.8	10.4	42.2
ByteDance_pr	1-2	13.7	10.1	42.0
Tencent_pr	3-4	11.6	8.6	39.4
Tencent_contr	3-4	11.5	8.8	38.9
GMU_pr	5-6	10.6	8.1	39.8
GMU_contr	5-6	10.5	8.0	39.4
DENTRA_pr	7	4.4	2.3	26.0

Table 100: Results in lin-nya, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	17.7	13.6	43.1
ByteDance_pr	1-2	17.6	13.6	42.9
Tencent_pr	3-5	11.2	9.0	34.8
GMU_contr	3-5	11.0	8.4	36.9
GMU_pr	3-5	10.5	7.9	36.5
Tencent_contr	6	10.4	8.8	33.8
DENTRA_pr	7	4.0	2.4	22.4

Table 101: Results in lin-kin, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	12.1	7.4	42.0
ByteDance_pr	1-2	12.0	7.2	42.0
Tencent_contr	3-4	7.1	4.8	35.2
Tencent_pr	3-4	7.1	5.0	34.8
GMU_pr	5-7	3.2	2.0	22.4
DENTRA_pr	5-7	3.2	1.8	23.9
GMU_contr	5-7	3.2	1.9	22.5

Table 102: Results in kin-lug, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	23.0	18.6	48.4
ByteDance_pr	1-2	23.0	18.6	48.6
GMU_pr	3-5	20.7	17.2	47.8
GMU_contr	3-5	20.7	17.2	47.8
Tencent_contr	3-5	20.5	17.1	47.1
Tencent_pr	6	20.3	16.9	47.1
DENTRA_pr	7	4.7	3.1	26.4

Table 103: Results in nya-swh, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	18.6	9.0	47.6
ByteDance_pr	1-2	18.6	9.1	47.6
GMU_contr	3-4	15.8	7.3	46.4
GMU_pr	3-4	15.7	7.4	46.3
Tencent_pr	5-6	14.8	6.8	45.7
Tencent_contr	5-6	14.7	6.8	45.3
DENTRA_pr	7	7.1	3.5	35.5

Table 104: Results in amh-zul, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	17.9	14.7	43.4
ByteDance_contr	1-2	17.7	14.5	43.3
Tencent_contr	3-5	14.1	11.6	40.4
GMU_contr	3-5	14.0	11.6	40.2
Tencent_pr	3-5	14.0	11.6	40.5
GMU_pr	6	13.7	11.5	39.7
DENTRA_pr	7	6.7	5.0	29.3

Table 105: Results in yor-swh, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	11.1	4.9	26.9
ByteDance_contr	2	10.9	4.8	26.6
Tencent_pr	3-4	5.2	3.2	21.8
Tencent_contr	3-4	5.0	3.0	21.5
GMU_contr	5	3.9	2.8	20.9
GMU_pr	6	3.7	2.7	20.8
DENTRA_pr	7	2.3	1.5	14.8

Table 106: Results in swh-yor, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	21.2	7.1	32.3
ByteDance_pr	1-2	21.1	7.2	32.2
GMU_pr	3	16.6	5.1	29.3
GMU_contr	4-5	16.4	5.0	29.3
Tencent_pr	4-5	16.4	5.3	28.8
Tencent_contr	6	15.5	4.9	28.1
DENTRA_pr	7	3.1	1.2	11.4

Table 107: Results in zul-amh, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	22.9	19.9	48.0
ByteDance_contr	2	22.6	19.6	47.9
Tencent_pr	3-4	20.0	17.5	45.3
Tencent_contr	3-4	19.8	17.1	45.9
GMU_pr	5-6	18.9	16.7	44.0
GMU_contr	5-6	18.7	16.5	43.9
DENTRA_pr	7	4.2	2.6	22.7

Table 108: Results in kin-hau, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	21.5	16.1	47.1
ByteDance_pr	1-2	21.4	16.0	47.1
Tencent_contr	3-4	18.0	13.6	44.0
Tencent_pr	3-4	17.8	13.3	43.5
GMU_pr	5	14.3	11.1	40.8
GMU_contr	6	14.2	10.9	40.6
DENTRA_pr	7	3.7	1.9	20.2

Table 109: Results in hau-kin, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	13.9	9.1	40.9
ByteDance_contr	2	13.7	9.1	40.7
GMU_contr	3	12.6	8.1	40.3
GMU_pr	4-6	12.4	8.0	40.3
Tencent_pr	4-6	12.4	8.0	40.5
Tencent_contr	4-6	12.3	8.1	40.0
DENTRA_pr	7	5.6	4.0	27.5

Table 110: Results in nya-som, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_contr	1-2	14.4	10.7	43.1
ByteDance_pr	1-2	14.4	10.7	43.4
Tencent_contr	3	13.3	10.1	42.5
Tencent_pr	4	13.1	9.9	42.4
GMU_pr	5-6	12.8	9.7	42.5
GMU_contr	5-6	12.8	9.7	42.4
DENTRA_pr	7	6.2	4.6	30.3

Table 111: Results in som-nya, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	12.1	7.3	41.2
ByteDance_contr	1-2	12.0	7.3	41.2
Tencent_pr	3	7.3	5.2	35.2
Tencent_contr	4	7.0	5.2	34.7
GMU_pr	5-6	5.7	4.4	31.4
GMU_contr	5-6	5.5	4.2	31.0
DENTRA_pr	7	2.3	1.5	23.2

Table 112: Results in xho-lug, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1-2	13.0	6.1	40.0
ByteDance_contr	1-2	12.9	6.1	39.8
Tencent_contr	3-4	11.2	5.0	37.6
Tencent_pr	3-4	11.1	5.1	37.8
GMU_pr	5-6	9.6	4.6	36.0
GMU_contr	5-6	9.5	4.7	36.0
DENTRA_pr	7	2.3	1.4	22.9

Table 113: Results in lug-xho, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	16.5	13.5	40.1
ByteDance_contr	2	15.9	13.0	39.6
Tencent_contr	3-4	11.7	9.0	36.3
Tencent_pr	3-4	11.6	9.1	36.5
GMU_contr	5	8.7	6.9	31.1
GMU_pr	6	8.3	6.6	30.0
DENTRA_pr	7	5.6	4.3	26.1

Table 114: Results in wol-swh, sorted by spBLEU.

System	Rank	spBLEU	BLEU	chrF2
ByteDance_pr	1	8.5	5.9	28.5
ByteDance_contr	2	8.3	5.9	28.0
GMU_pr	3-5	3.5	2.5	15.9
GMU_contr	3-5	3.4	2.3	15.0
Tencent_pr	3-5	3.4	2.7	19.1
DENTRA_pr	6-7	3.0	1.8	18.8
Tencent_contr	6-7	3.0	2.4	18.0

Table 115: Results in swh-wol, sorted by spBLEU.