Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages

∀*, Wilhelmina Nekoto[1], Vukosi Marivate[2], Tshinondiwa Matsila[1], Timi Fasubaa[3], Tajudeen Kolawole[4], Taiwo Fagbohungbe[5], Solomon Oluwole Akinola[6], Shamsuddee Hassan Muhammad[7,39], Salomon Kabongo[4], Salomey Osei[4], Sackey Freshia[8], Rubungo Andre Niyongabo[9], Ricky Macharm[10], Perez Ogayo[11], Orevaoghene Ahia[12], Musie Meressa[13], Mofe Adeyemi[14], Masabata Mokgesi-Selinga[15], Lawrence Okegbemi[5], Laura Jane Martinus[16], Kolawole Tajudeen[4], Kevin Degila[17], Kelechi Ogueji[12], Kathleen Siminyu[18], Julia Kreutzer[19], Jason Webster[20], Jamiil Toure Ali[1], Jade Abbott[21], Iroro Orife[3], Ignatius Ezeani[38], Idris Abdulkabir Dangana[23,7], Herman Kamper[24], Hady Elsahar[25], Goodness Duru[26], Ghollah Kioko[27], Espoir Murhabazi[1], Elan van Biljon[12,24], Daniel Whitenack[28], Christopher Onyefuluchi[29], Chris Emezue[40], Bonaventure Dossou[31], Blessing Sibanda[32], Blessing Itoro Bassey[4], Ayodele Olabiyi[33], Arshath Ramkilowan[34], Alp Öktem[35], Adewale Akinfaderin[36], Abdallah Bashir[37]

*Masakhane, Africa [1]Independent, [2]University of Pretoria, [3]Niger-Volta LTI,
[4]African Masters in Machine Intelligence, [5]Federal University of Technology, Akure, [6]University of Johannesburg,
[7]Bayero University, Kano, [8]Jomo Kenyatta University of Agriculture and Technology, [9]UESTC,
[10]Siseng Consulting Ltd, [11]African Leadership University, [12]InstaDeep Ltd,
[13]Sapienza University of Rome, [14]Udacity, [15]Parliament, Republic of South Africa,
[16]Explore Data Science Academy, [17]UCD, Konta, [18]AI4Dev, [19]Google Research, [20]Percept, [21]Retro Rabbit,
[23]Di-Hub, [24]Stellenbosch University, [25]Naver Labs Europe, [26]Retina, AI, [27]Lori Systems, [28]SIL International,
[29]Federal College of Dental Technology and Therapy, Enugu, [31]Jacobs University,
[32]Namibia University of Science and Technology, [33]Data Science Nigeria, [34]Praekelt Consulting,
[35]Translators without Borders, [36]Amazon, [37]Max Planck Institute for Informartics, University of Saarland,
[38]Lancaster University, [39]University of Porto, [40]Technical University of Munich

masakhane-nlp@googlegroups.com

Abstract

Research in NLP lacks geographic diversity, and the question of how NLP can be scaled to low-resourced languages has not yet been adequately solved. "Low-resourced"-ness is a complex problem going beyond data availability and reflects systemic problems in society.

In this paper, we focus on the task of Machine Translation (MT), that plays a crucial role for information accessibility and communication worldwide. Despite immense improvements in MT over the past decade, MT is centered around a few high-resourced languages. As MT researchers cannot solve the problem of low-resourcedness alone, we propose participatory research as a means to involve all necessary agents required in the MT development process. We demonstrate the feasibility and scalability of participatory research with a case study on MT for African languages. Its implementation leads to a collection of novel translation datasets, MT benchmarks for over 30 languages, with human evaluations for a third of them, and enables participants without formal training to make a unique scientific contribution. Benchmarks, models, data, code, and evaluation results are released at https://github.com/masakhane-io/masakhane-mt.

∀ to represent the whole Masakhane community.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2144–2160
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

1 Introduction

Language prevalence in societies is directly bound to the people and places that speak this language. Consequently, resource-scarce languages in an NLP context reflect the resource scarcity in the society from which the speakers originate (McCarthy, 2017). Through the lens of a machine learning researcher, "low-resourced" identifies languages for which few digital or computational data resources exist, often classified in comparison to another language (Gu et al., 2018; Zoph et al., 2016). However, to the sociolinguist, "low-resourced" can be broken down into many categories: low density, less commonly taught, or endangered, each carrying slightly different meanings (Cieri et al., 2016). In this complex definition, the "low-resourced"-ness of a language is a symptom of a range of societal problems, e.g. authors oppressed by colonial governments have been imprisoned for writing novels in their languages impacting the publications in those languages (Wa Thiong'o, 1992), or that fewer PhD candidates come from oppressed societies due to low access to tertiary education (Jowi et al., 2018). This results in fewer linguistic resources and researchers from those regions to work on NLP for their language. Therefore, the problem of "low-resourced"-ness relates not only to the available resources for a language, but also to the *lack of geographic and language diversity of NLP researchers* themselves.

The NLP community has awakened to the fact that it has a diversity crisis in terms of limited geographies and languages (Caines, 2019; Joshi et al., 2020): Research groups are extending NLP research to low-resourced languages (Guzmán et al., 2019; Hu et al., 2020; Wu and Dredze, 2020), and workshops have been established (Haffari et al., 2018; Axelrod et al., 2019; Cherry et al., 2019).

We scope the rest of this study to machine translation (MT) using parallel corpora only, and refer the reader to Joshi et al. (2019) for an assessment of low-resourced NLP in general.

Language	Articles	Speakers	Category
English	6,087,118	1,268,100,000	Winner
Egyptian Arabic	573,355	64,600,000	Hopeful
Afrikaans	91,002	17,500,000	Rising Star
Kiswahili	59,038	98,300,000	Rising Star
Yoruba	32,572	39,800,000	Rising Star
Shona	5,505	9,000,000	Scraping by
Zulu	2,219	27,800,000	Hopeful
Igbo	1,487	27,000,000	Scraping by
Luo	0	4,200,000	Left-behind
Fon	0	2,200,000	Left-behind
Dendi	0	257,000	Left-behind
Damara	0	200,000	Left-behind

Table 1: Sizes of a subset of African language Wikipedias[1], speaker populations[2], and categories according to Joshi et al. (2020) (28 May 2020).

Contributions. We diagnose the problems of MT systems for low-resourced languages by reflecting on what agents and interactions are necessary for a sustainable MT research process. We identify which agents and interactions are commonly omitted from existing low-resourced MT research, and assess the impact that their exclusion has on the research. To involve the necessary agents and facilitate required interactions, we propose *participatory research to build sustainable MT research communities for low-resourced languages*. The feasibility and scalability of this method is demonstrated with a case study on MT for African languages, where we present its implementation and outcomes, including novel translation datasets, benchmarks for over 30 target languages contributed and evaluated by language speakers, and publications authored by participants without formal training as scientists.

2 Background

Cross-lingual Transfer. With the success of deep learning in NLP, language-specific feature design has become rare, and cross-lingual transfer methods have come into bloom (Upadhyay et al., 2016; Ruder et al., 2019) to transfer progress from high-resourced to low-resourced languages (Adams et al., 2017; Wang et al., 2019; Kim et al., 2019). The most diverse benchmark for multilingual transfer by Hu et al. (2020) allows measurement of the success of such transfer approaches across 40 languages from 12 language families. However, the inclusion of languages in the set of benchmarks is dependent on the availability of monolingual data for representation learning with previously annotated resources. The content of the benchmark tasks is English-sourced, and human performance estimates are taken from English. Most cross-lingual representation learning techniques are Anglo-centric in their design (Anastasopoulos and Neubig, 2019).

Multilingual Approaches. Multilingual MT (Dong et al., 2015; Firat et al., 2016a,b; Wang et al., 2020) addresses the transfer of MT from high-resourced to low-resourced languages by training multilingual models for all languages at once. (Aharoni et al., 2019; Arivazhagan et al., 2019) train models to translate between English and 102 languages, for the 10 most high-resourced African languages on private data, and otherwise on public TED talks (Qi et al., 2018). Multilingual training often outperforms bilingual training, especially for low-resourced languages. However, with multilingual parallel data being also Anglo-centric, the capabilities to translate from English versus into English vastly diverge (Zhang et al., 2020).

Another recent approach, mBART (Liu et al., 2020), leverages both monolingual and parallel data and also yields improvements in translation quality for lower-resource languages such as Nepali, Sinhala and Gujarati.[3] While this provides a solution for small quantities of training data or monolingual resources, the extent to which standard BLEU evaluations reflect translation quality is not clear yet, since human evaluation studies are missing.

Targeted Resource Creation. Guzmán et al. (2019) develop evaluation datasets for low-resourced MT between English and Nepali, Sinhala, Khmer and Pashtolow. They highlight many problems with low-resourced translation: tokenization, content selection, and translation verification, illustrating increased difficulty translating from English into low-resourced languages, and highlight the ineffectiveness of accepted state-of-the-art techniques on morphologically-rich languages. Despite involving all agents of the MT process (Section 3), the study does not involve data curators or evaluators that understood the languages involved, and resorts to standard MT evaluation metrics. Additionally, how this effort-intensive approach would scale to more than a handful of languages remains an open question.

3 The Machine Translation Process

We reflect on the process enabling a sustainable process for MT research on parallel corpora in terms of the required agents and interactions, visualized in Figure 1. Content creators, translators, and curators form the dataset creation process, while the language technologists and evaluators are part of the model creation process. Stakeholders (not displayed) create demand for both processes.

Stakeholders are people impacted by the artifacts generated by each agent in the MT process, and can typically speak and read the source or the target languages. To benefit from MT systems, the stakeholders need access to technology and electricity.

Content Creators produce content in a language, where content is any digital or non-digital representation of language. For digi-

[3] Note that these languages have more digital resources available and a longer history of written texts than the low-resourced languages we are addressing here.

tal content, content creators require keyboards, and access to technology.

Translators translate the original content, including crowd-workers, researchers, or translation professionals. They must understand the language of the content creator and the target language. A translator needs content to translate, provided by content creators. For digital content, the translator requires keyboards and technology access.

Curators are defined as individuals involved in the content selection for a dataset (Bender and Friedman, 2018), requiring access to content and translations. They should understand the languages in question for quality control and encoding information.

Language Technologists are defined as individuals using datasets and computational linguistic techniques to produce MT models between language pairs. Language technologists require language preprocessors, MT toolkits, and access to compute resources.

Evaluators are individuals who measure and analyse the performance of a MT model, and therefore need knowledge of both source and target languages. To report on the performance on models, evaluators require quality metrics, as well as evaluation datasets. Evaluators provide feedback to the Language Technologists for improvement.

3.1 Limitations of Existing Approaches

If we place a high-resource MT pair such as English-to-French into the process defined above, we observe that each agent nowadays has the necessary resources and historical stakeholder demand to perform their role effectively. A "virtuous cycle" emerged where available content enabled the development of MT systems that in turn drove more translations, more tools, more evaluation and more content, which cycled back to improving MT systems.

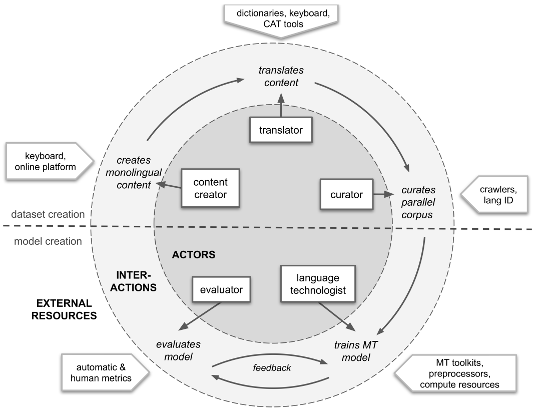


Figure 1: The MT Process, in terms of the necessary agents, interactions and external constraints and demand (excluding stakeholders).

By contrast, parts of the process for existing low-resourced MT are constrained. Historically, many low-resourced languages had *low demand from stakeholders* for content creation and translation (Wa Thiong'o, 1992). Due to *missing keyboards or limited access to technology*, content creators were not empowered to write digital content (Adam, 1997; van Esch et al., 2019). This is a chicken-or-egg problem, where existing digital content in a language would attract more stakeholders, which would incentivize content creators (Kaffee et al., 2018). As a result, primary data sources for NLP research, such as Wikipedia, often have a few hundred articles only for low-resourced languages despite large speaker populations, see Table 1. Due to limited demand, existing translations are often domain-specific and small in size, such as the JW300 corpus (Agić and Vulić, 2019) whose content was created for missionary purposes.

When data curators are not part of the societies from where these languages originate, they are are often unable to identify data sources or translators for languages, prohibiting them from checking the *validity of the created resource*. This creates problems in encoding, orthography or alignment, resulting in noisy or incorrect translation pairs (Taylor et al., 2015). This is aggravated by the fact that many low-resourced languages do not have a long written history to draw from and therefore might be less standardized and using multiple scripts. In collaboration with content creators, data curators can contribute to *standardization* or at least recognize potential issues for data processing further down the line.

As discussed in Section 1, language technologists are fewer in low-resourced societies. Furthermore, the *techniques developed in high-resourced societies might be inapplicable* due to compute, infrastructure or time constraints. Aside from the problem of education and complexity, existing techniques may not apply due to linguistic and morphological differences in the languages, or the scale, domain, or quality of the data (Hu et al., 2020; Pires et al., 2019).

Evaluators usually resort to potentially *unsuitable automatic metrics* due to time constraints or missing connections to stakeholders (Guzmán et al., 2019). The main evaluators of low-resourced NLP that is developed today typically cannot use human metrics due to the inability to speak the languages, or the lack of reliable crowdsourcing infrastructure, identified as one of the core weaknesses of previous approaches (in Section 2).

In summary, many agents in the MT process for low-resourced languages are either missing invaluable language and societal knowledge, or the necessary technical resources, knowledge, connections, and incentives to form interactions with other agents in the process.

3.2 Participatory Research Approach

We propose one way to overcome the limitations in Section 3.1: ensuring that *the agents in the MT process originate from the countries where the low-resourced languages are spoken* or can speak the low-resourced languages. Where this condition cannot be satisfied, at least a *knowledge transfer* between agents should be enabled. We hypothesize that using a participatory approach will allow researchers to improve the MT process by iterating faster and more effectively.

Participatory research, unlike conventional research, emphasizes the value of research partners in the knowledge-production process where the research process itself is defined collaboratively and iteratively. The "participants" are individuals involved in conducting research without formal training as researchers. Participatory research describes a broad set of methodologies, organised in terms of the level of participation. At the lowest level is crowd-sourcing, where participants are involved solely in data collection. The highest level—extreme citizen science–involves participation in the problem definition, data collection, analysis and interpretation (English et al., 2018).

Crowd-sourcing has been applied to low-resourced language data collection (Ambati et al., 2010; Guevara-Rukoz et al., 2020; Millour and Fort, 2018), but existing studies highlight how the disconnect between the data creation process and model creation process causes challenges. In seeking to create cross-disciplinary teams that emphasize the values in a societal context, a participatory approach which involves participants in every part of the scientific process appears pertinent to solving the problems for low-resourced languages highlighted in Section 3.1.

To show how more involved participatory research can benefit low-resource language translation, we present a case study in MT for African languages.

4 Case Study: Masakhane

Africa currently has 2144 living languages (Eberhard et al., 2019). Despite this, African languages account for a small fraction of available language resources, and NLP research rarely considers African languages. In the taxonomy of Joshi et al. (2020), African languages are assigned categories ranging from "The Left Behinds" to "The Rising Stars", with most languages not having any annotated data. Even monolingual resources are sparse, as shown in Table 1.

In addition to a lack of NLP datasets, the African continent lacks NLP researchers. In 2018, only five out of the 2695 affiliations of the five major NLP conferences were from African institutions (Caines, 2019). ∀ et al. (2020) attribute this to a culmination of circumstances, in particular their societal embedding (Alexander, 2009) and socio-economic factors, hindering participation in research activities and events, leaving researchers disconnected and distributed across the continent. Consequently, existing data resources are harder to discover, especially since these are often published in closed journals or are not digitized (Mesthrie, 1995).

For African languages, the implementation of a standard crowd-sourcing pipeline as for example used for collecting task annotations for English, is at the current stage infeasible, due to the challenges outlined in Section 3 and above. Additionally, no standard MT evaluation set for all of the languages in focus exists, nor are there prior published systems that we could compare all models against for a more insightful human evaluation. We therefore resort to intrinsic evaluation, and rely on this work becoming the *first benchmark for future evaluations*.

We invite the reader to adopt a meta-perspective of this case study as an empirical experiment: Where the *hypothesis* is that participatory research can facilitate low-resourced MT development; the *experimental methodology* is the strategies and tools employed to bring together distributed participants, enabling each language speaker to train, contribute, and evaluate their models. The experiment is *evaluated* in terms of the quantity and diversity of participants and languages, and the variety of research artifacts, in terms of benchmarks, human evaluations, publications, and the overall health of the community. While a set of novel human evaluation results are presented, they serve as demonstration of the value of a participatory approach, rather than the empirical focus of the paper.

4.1 Methodology

To overcome the challenge of recruiting participants, a number of strategies were employed. Starting from local demand at a machine learning school (Deep Learning Indaba (Engelbrecht, 2018)), meetups and universities, distant connections were made through Twitter, conference workshops,[4] and eventually press coverage[5] and research publications.[6] To overcome the limited tertiary education enrollments in Sub-Saharan Africa (Jowi et al., 2018), *no prerequisites* were placed on researchers joining the project. For the agents outlined in Section 3, no fixed roles are imposed onto participants. Instead, they join with a specific interest, background, or skill aligning them best to one or more of agents. To obtain cross-disciplinarity, we focus on the communication and interaction between participants to enable knowledge transfer between missing connections (identified in Section 3.1), allowing a fluidity of agent roles. For example, someone who initially joined with the interest of using

[4] ICLR AfricaNLP 2020: https://africanlp-workshop.github.io/

[5] https://venturebeat.com/2019/11/27/the-masakhane-project-wants-machine-translation-and-ai-to-transform-africa/

[6] https://github.com/masakhane-io/masakhane-community/blob/master/publications.md

machine translation for their local language (as a stakeholder) to translate education material, might turn into a junior language technologist when equipped with tools and introductory material and mentoring, and guide content creation more specifically for resources needed for MT.

To bridge large geographical divides, the community lives online. Communication occurs on GitHub and Slack with weekly video conference meetings and reading groups. Meeting notes are shared openly so that continuous participation is not required and time commitment can be organized individually. Sub-interest groups have emerged in Slack channels to allow focused discussions. Agendas for meetings and reading groups are public and democratically voted upon. In this way, the research questions evolve based on *stakeholder demands*, rather than being imposed upon by external forces.

The lack of compute resources and prior exposure to NLP is overcome by providing tutorials for training a custom-size Transformer model with JoeyNMT (Kreutzer et al., 2019) on Google Colab[7]. International researchers were not prohibited from joining. As a result, mutual mentorship relations emerged, whereby international researchers with more language technology experience guided research efforts and enabled data curators or translators to become language technologists. In return, African researchers introduced the international language technologists to African stakeholders, languages and context.

4.2 Research Outcomes

Participants. A growth to over 400 participants of diverse disciplines, from at least 20 countries, has been achieved within the past year, suggesting the participant recruitment process was effective. Appendix A contains detailed demographics of a subset of participants from a voluntary survey in February 2020. 86.5% of participants responded positively when asked if the community helped them find mentors or collaborators, indicating that the health of the community is positive. This is also reflected in joint research publications of new groups of collaborators.

Research Artifacts. As a result of mentorship and knowledge exchange between agents of the translation process, our implementation of participatory research has produced artifacts for NLP research, namely datasets, benchmarks and models, which are publicly available online.[8]. Additionally, over 10 participants have gone on to publish works addressing language-specific challenges at conference workshops, such as (Dossou and Emezue, 2020; Orife, 2020; Orife et al., 2020; Öktem et al., 2020; Van Biljon et al., 2020; Martinus et al., 2020; Marivate et al., 2020).

Dataset Creation. The dataset creation process is ongoing, with new initiatives still emerging. We showcase a few initiatives below to demonstrate how bridging connections between agents facilitates the MT process.

1. A team of Nigerian participants, driven by the internal demand to ensure that accessible and representative data of their culture is used to train models, are translating their own writings including personal religious stories and undergraduate theses into Yoruba and Igbo[9].

2. A Namibian participant, driven by a passion to preserve the culture of the Damara, is hosting collaborative sessions with Damara speakers, to collect and translate phrases that reflect Damara culture

[7] https://colab.research.google.com

[8] https://github.com/masakhane-io

[9] https://github.com/masakhane-io/masakhane-wazobia-dataset

around traditional clothing, songs, and prayers.[10]

3. Creating a connection between a translator in South-Africa's parliament and a language technologist has enabled the process of data curation, allowing access to data from the parliament in South-Africa's languages (which are public but obfuscated behind internal tools).[11].

These stories demonstrate the value of including curators, content creators, and translators as participants.

Benchmarks. We publish 45 benchmarks for neural translation models from English into 32 distinct African languages, and from French into two additional languages, as well as from English into three different languages.[12] Most were trained on the JW300 corpus (Agić and Vulić, 2019). From this corpus, we select the English sentences most commonly found (and longer than 4 tokens) in all languages, as a global set of test sources. For individual languages, test splits are composed by selecting the translations that are available from this subset. While this biases the test set towards frequent segments, it prevents cross-lingual overlap between training and test data which has to be ensured for cross-lingual transfer learning. For training data, other sources like Autshumato (McKellar, 2014), TED (Cettolo et al., 2012), SAWA (De Pauw et al., 2009), Tatoeba[13], Opus (Tiedemann, 2012), and data translated or curated by participants were added. Language pairs were selected based on the individual demands of each of the 32 participants, who voluntarily contributed the benchmarks they valued most. 16 of the selected target languages are categorized as "Left-behind" and 11 are categorized as "Scraping by" in the taxonomy of (Joshi et al., 2020). The benchmarks are hosted publicly, including model weights, configurations and preprocessing pipelines for full reproducibility. The benchmarks are submitted by individual or groups of participants in form of a GitHub Pull Request. By this, we ensure that the contact to the benchmark contributors can be made, and ownership is experienced.

4.3 Human MT Evaluation

To our knowledge, there is no prior research on human evaluation specifically for machine translations of low-resourced languages. Until now, NLP practitioners were left with the hope that successful evaluation methodologies for high-resource languages would transfer well to low-resourced languages. This lack of study is due to the missing connections between the community of speakers (content creators and translators), and the language technologists. MT evaluations by humans are often done either within a group of researchers from the same lab or field (e.g. for WMT evaluations[14]), or via crowdsourcing platforms (Ambati and Vogel, 2010; Post et al., 2012). Speakers of low-resource languages are traditionally underrepresented in these groups, which makes such studies even harder (Joshi et al., 2019; Guzmán et al., 2019).

One might argue that human evaluation should not be attempted before reaching a viable state of quality, but we found that early evaluation results in an improved understanding of the individual challenges of the target languages, strengthens the network of the community, and most importantly, improves the connection and knowledge transfer between language technologists, content creators and

[10] https://github.com/masakhane-io/masakhane-khoekhoegowab

[11] http://bit.ly/raw-parliamentary-translations

[12] Benchmark scores can be found in Appendix C.

[13] https://tatoeba.org/

[14] http://www.statmt.org/wmt19/

curators.

The "low-resourced"-ness of the addressed languages pose challenges for evaluation beyond interface design or recruitment of evaluators proficient in the target language. For the example of Igbo, evaluators had to find solutions for typing diacritics without a suitable keyboard. In addition, Igbo has many dialects and variations which the MT model is uninformed of. Medical or technical terminology (e.g., "data") is difficult to translate and whether to use loan words required discussion. Target language news websites were found to be useful for resolving standardization or terminology questions. Solutions for each language were shared and often also applicable for other languages.

Data. The models are trained on JW300 data.[15] To gain real-world quality estimates beyond religious context, we assess the models' out-of-domain generalization by translating a English COVID-19 survey with 39 questions and statements regarding COVID-19,[16] where the human-corrected and approved translations can directly serve the purpose of gathering responses. The domain is challenging as it contains medical terms and new vocabulary. Furthermore, we evaluate a subset of the Multitarget TED test data (Duh, 2018)[17]. The obtained translations enrich the TED datasets, adding new languages for which no prior translations exist. The size of the TED evaluations vary from 30 to 120 sentences. Details are given in Table 3, Appendix B.

Evaluators. 11 participants of the community volunteered to evaluate translations in their language(s), often involving family or friends to determine the most correct translations. The evaluator role is therefore taken by both stakeholders and language technologists. Within only 10 days, we gathered a total of 707 evaluated translations covering Igbo (*ig*), Nigerian Pidgin (*pcm*), Shona (*sn*), Luo (*luo*), Hausa (*ha*, twice by two different annotators), Kiswahili (*sw*), Yoruba (*yo*), Fon (*fon*) and Dendi (*ddn*). We did not impose prescriptions in terms of number of sentences to evaluate, or time to spend, since this was voluntary work, and guidelines or estimates for the evaluation of translations into these languages are non-existent.

Evaluation Technique. Instead of a direct assessment (Graham et al., 2013) often used in benchmark MT evaluations (Barrault et al., 2019; Guzmán et al., 2019), we opt for post-editing. Post-edits are grounded in actions that can be analyzed in terms of e.g. error types for further investigations, while direct assessments require expensive calibration (Bentivogli et al., 2018). Embedded in the community, these post-edit evaluations create an asset for the interaction of various agents: for the language technologists for domain adaptation, or for the content creators, curators, or translators for guidance in standardization or domain choice.

Results. Table 2 reports evaluation results in terms of BLEU evaluated on the benchmark test set from JW300, and human-targeted TER (HTER) (Snover et al., 2006), BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) against human corrected model translations. For *ha* we find modest agreement between evaluators: Spearman's $\rho = 0.56$ for sentence-BLEU measurements of the post-edits compared to the original hypotheses. Generally, we observe that the JW300 score is misleading, overestimating model quality (except *yo*). Training data size appears to be a more reliable predictor of generalization abilities, illustrating the danger of chasing a single benchmark. However, *ig* and *yo* both have comparable amounts

[15] Except for Hausa: multiple domains, see Table 4.

[16] https://coronasurveys.org/

[17] http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/

Trg.	Train.	Autom.: JW300	Human: COVID			Human: TED		
lang.	size	BLEU ↑	HTER ↓	HBLEU ↑	HCHRF ↑	HTER ↓	HBLEU ↑	HCHRF ↑
ddn	6,937	22.30	1.11	0.27	0.08	-	-	-
pcm	20,214	23.29	0.98	3.03	0.19	0.84	9.76	25.16
fon	27,510	31.07	0.92	15.43	23.22	-	-	-
luo	136,459	34.33	-	-	-	1.26	7.90	20.88
ha	333,845	41.11	0.71	26.96	43.97	0.73	20.42	39.31
			0.64	26.56	46.71	-	-	-
ig	414,467	34.85	0.85	11.94	29.86	0.55	33.74	49.67
yo	415,100	38.62	0.09	85.92	89.90	0.51	49.22	58.41
sn	712,455	30.84	0.53	31.31	54.04	-	-	-
sw	875,558	48.94	-	-	-	0.32	60.47	78.67

Table 2: Evaluation results for translations from English. Metrics are computed based on Polyglot-tokenized translations. HTER are mean sentence-level TER scores computed with the Pyter Python package. BLEU and ChrF are computed with Sacrebleu and tokenize "none" (Post, 2018).

of training data, JW300 scores, and carry diacritics, but exhibit very different evaluation performances, in particular on COVID. This can be explained by the large variations of *ig* as discussed above: Training data and model output are not consistent with respect to one dialect, while the *evaluator* had to decide on one. We also find difference in performance across domains, with the TED domain appearing easier for *pcm* and *ig*, while the *yo* model performs better on COVID.

5 Conclusion

We proposed a participatory approach as a solution to sustainably scaling NLP research to low-resourced languages. Having identified key agents and interactions in the MT development process, we implement a participatory approach to build a community for African MT. In the process, we discovered successful strategies for distributed growth and communication, knowledge sharing and model building. In addition to publishing benchmarks and datasets for previously understudied languages, we show how the participatory design of the community enables us to conduct a human evaluation study of model outputs, which has been one of the limitations of previous approaches to low-resourced NLP. The sheer volume and diversity of participants, languages and outcomes, and that for many for languages featured, this paper constitutes the first time that human evaluation of an MT system has been performed, is evidence of the value of participatory approaches for low-resourced MT. For future work, we will (1) continue to iterate, analyze and widen our benchmarks and evaluations, (2) build richer and more meaningful datasets that reflect priorities of the stakeholders, (3) expand the focus of the existing community for African languages to other NLP tasks, and (4) help implement similar communities for other geographic regions with low-resourced languages.

Acknowledgements

We would like to thank Benjamin Rosman and Richard Klein for their invaluable feedback, as well as the anonymous EMNLP reviewers, and George Foster and Daan van Esch. We would also like to thank Google Cloud for the grant that enabled us to build baselines for languages with larger datasets.

References

Lishan Adam. 1997. Content and the web for african development. *Journal of information science*, 23(1):91–97.

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947, Valencia, Spain. Association for Computational Linguistics.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Neville Alexander. 2009. Evolving african approaches to the management of linguistic diversity: The acalan project. *Language Matters*, 40(2):117–132.

Vamshi Ambati and Stephan Vogel. 2010. Can crowds build parallel corpora for machine translation systems? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 62–65, Los Angeles. Association for Computational Linguistics.

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2010. Active learning-based elicitation for semi-supervised word alignment. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 365–370, Uppsala, Sweden. Association for Computational Linguistics.

Antonios Anastasopoulos and Graham Neubig. 2019. Should all cross-lingual embeddings speak english?

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Amittai Axelrod, Diyi Yang, Rossana Cunha, Samira Shaikh, and Zeerak Waseem, editors. 2019. *Proceedings of the 2019 Workshop on Widening NLP*. Association for Computational Linguistics, Florence, Italy.

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. 2019. 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. Association for Computational Linguistics.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Luisa Bentivogli, Mauro Cettolo, Marcello Federico, and Federmann Christian. 2018. Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment. In *15th International Workshop on Spoken Language Translation 2018*, pages 62–69.

Andrew Caines. 2019. The geographic diversity of nlp conferences.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Colin Cherry, Greg Durrett, George Foster, Reza Haffari, Shahram Khadivi, Nanyun Peng, Xiang Ren, and Swabha Swayamdipta, editors. 2019. *Proceedings of the 2nd Workshop on*

Deep Learning Approaches for Low-Resource NLP (DeepLo 2019). Association for Computational Linguistics, Hong Kong, China.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4543–4549, Portorož, Slovenia. European Language Resources Association (ELRA).

Guy De Pauw, Peter Waiganjo Wagacha, and Gilles-Maurice de Schryver. 2009. The SAWA corpus: A parallel corpus English - Swahili. In *Proceedings of the First Workshop on Language Technologies for African Languages*, pages 9–16, Athens, Greece. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Bonaventure FP Dossou and Chris C Emezue. 2020. Ffr v1. 0: Fon-french neural machine translation. *arXiv preprint arXiv:2003.12111*.

Kevin Duh. 2018. The multitarget ted talks task. `http://www.cs.jhu.edu/~kevinduh/a/multitarget-tedtalks/`.

David M Eberhard, Gary F. Simons, and Charles D. Fenning. 2019. Ethnologue: Languages of the worlds. twenty-second edition.

Herman Engelbrecht. 2018. The deep learning indaba report. *ACM SIGMultimedia Records*, 9(3):5–5.

PB English, MJ Richardson, and Catalina Garzón-Galvis. 2018. From crowdsourcing to extreme citizen science: participatory research for environmental health. *Annual review of public health*, 39:335–350.

Daan van Esch, Elnaz Sarbar, Tamar Lucassen, Jeremy O'Brien, Theresa Breiner, Manasa Prasad, Evan Crew, Chieu Nguyen, and Françoise Beaufays. 2019. Writing across the world's languages: Deep internationalization for gboard, the google keyboard. *arXiv preprint arXiv:1912.01218*.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

∀, Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade 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. In *"AfricaNLP" Workshop at the 8th International Conference on Learning Representations*.

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.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Volume 1 (Long Papers), pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Adriana Guevara-Rukoz, Isin Demirsahin, Fei He, Shan-Hui Cathy Chu, Supheakmungkol Sarin, Knot Pipatsrisawat, Alexander Gutkin, Alena Butryna, and Oddur Kjartansson. 2020. Crowdsourcing Latin American Spanish for low-resource text-to-speech. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6504–6513, Marseille, France. European Language Resources Association.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali–english and sinhala–english. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6100–6113.

Reza Haffari, Colin Cherry, George Foster, Shahram Khadivi, and Bahar Salehi, editors. 2018. *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*. Association for Computational Linguistics, Melbourne.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. In *Sixteenth International Conference on Natural Language Processing (ICON)*, Hyderabad, India.

Pratik M. Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *ArXiv*, abs/2004.09095.

James Jowi, Charles Ochieng Ong'ondo, and Mulu Nega. 2018. Building phd capacity in sub-saharan africa.

Lucie-Aimée Kaffee, Hady ElSahar, Pavlos Vougiouklis, Christophe Gravier, Frédérique Laforest, Jonathon Hare, and Elena Simperl. 2018. Mind the (language) gap: generation of multilingual wikipedia summaries from wikidata for articleplaceholders. In *European Semantic Web Conference*, pages 319–334. Springer.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, Hong Kong, China.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. *arXiv preprint arXiv:2003.04986*.

Laura Martinus, Jason Webster, Joanne Moonsamy, Moses Shaba Jnr, Ridha Moosa, and Robert Fairon. 2020. Neural machine translation for south africa's official languages. *arXiv preprint arXiv:2005.06609*.

Arya McCarthy. 2017. The new digital divide: Language is the impediment to information access. `https://hilltopicssmu.`

wordpress.com/2017/04/08/the-new-digital-divide-language-is-the-impediment-to-information-access/. Accessed: 2020-05-30.

Cindy A. McKellar. 2014. An english to xitsonga statistical machine translation system for the government domain. In *Proceedings of the 2014 PRASA, RobMech and AfLaT International Joint Symposium*, pages 229–233.

Rajend Mesthrie. 1995. *Language and social history: Studies in South African sociolinguistics*. New Africa Books.

Alice Millour and Karën Fort. 2018. Toward a lightweight solution for less-resourced languages: Creating a POS tagger for alsatian using voluntary crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Alp Öktem, Mirko Plitt, and Grace Tang. 2020. Tigrinya neural machine translation with transfer learning for humanitarian response. *arXiv preprint arXiv:2003.11523*.

Iroro Orife. 2020. Towards neural machine translation for edoid languages. *arXiv preprint arXiv:2003.10704*.

Iroro Orife, David I Adelani, Timi Fasubaa, Victor Williamson, Wuraola Fisayo Oyewusi, Olamilekan Wahab, and Kola Tubosun. 2020. Improving yor\ub\'a diacritic restoration. *arXiv preprint arXiv:2003.10564*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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

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

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada. Association for Computational Linguistics.

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

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Alex S Taylor, Siân Lindley, Tim Regan, David Sweeney, Vasillis Vlachokyriakos, Lillie Grainger, and Jessica Lingel. 2015. Data-in-place: Thinking through the relations between data and community. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2863–2872.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1661–1670, Berlin, Germany. Association for Computational Linguistics.

Elan Van Biljon, Arnu Pretorius, and Julia Kreutzer. 2020. On optimal transformer depth for low-resource language translation. *arXiv preprint arXiv:2004.04418*.

Ngugi Wa Thiong'o. 1992. *Decolonising the mind: The politics of language in African literature*. East African Publishers.

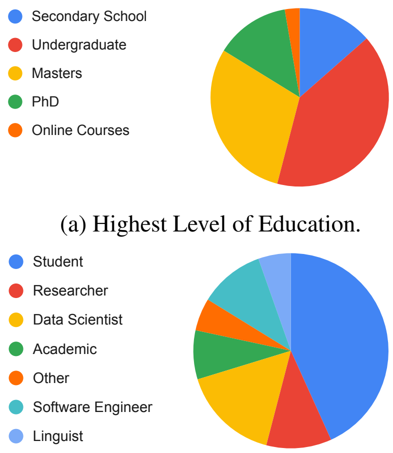
Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020. Balancing training for multilingual neural machine translation.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.



(a) Highest Level of Education.

(b) Occupation.

Figure 2: Education (a) and occupation (b) of a subset of 37 participants as indicated in a voluntary survey in February 2020.

Language	Domain	Size
Nigerian Pidgin	COVID	39
	TED	100
Luo	TED	30
Yoruba	COVID	39
	TED	80
Hausa	COVID	78
	TED	120
Igbo	COVID	39
	TED	50
Fon	COVID	39
Swahili	TED	55
Shona	COVID	39
Dendi	COVID	39

Table 3: Number of sentences for collected post-edits for TED talks and COVID surveys.

A Demographics

Figure 2 shows the demographics for a subset of participants from a voluntary survey conducted in February 2020. Between then and now (May 2020), the community has grown by 30%, so these figures have to be seen as a snapshot. Nevertheless we can see that the educational background and the occupation is fairly diverse, with a majority of undergraduate students (not necessarily Computer Science).

B Evaluation Data

Table 3 reports the number sentences that were post-edited in the human evaluation study reported in Section 4.

C Benchmark Scores

Table 4 contains BLEU scores on the JW300 test set for all benchmark models. BLEU scores are computed with Sacrebleu (Post, 2018) with tokenizer 'none' since the JW300 data comes tokenized with Polyglot.[18]. The table also features the target categories according to (Joshi et al., 2020) as of 28 May 2020.

[18] https://polyglot.readthedocs.io/en/latest/index.html

Source	Target	Best Test BLEU	Category
English	Afrikaans (Autshumato)	19.56	Rising Star
English	Afrikaans (JW300)	45.48	Rising Star
English	Amharic	2.03	Rising Star
English	Arabic (TED, custom)	9.28	Underdog
English	Dendi	22.30	Left Behind
English	Efik	33.48	Left Behind
English	Ẹ̀dó	12.49	Left Behind
English	Ẹ̀ṣ̀án	6.2	Left Behind
English	Fon	31.07	Left Behind
English	Hausa (JW300+Tatoeba+more)	41.11	Hopeful
English	Igbo	34.85	Scraping by
English	Isoko	38.91	Left Behind
English	Kamba	27.90	Left Behind
English	Kimbundu	32.76	Left Behind
English	Kikuyu	37.85	Scraping by
English	Lingala	48.64	Scraping by
English	Luo	34.33	Left Behind
English	Nigerian Pidgin	23.29	Left Behind
English	Northern Sotho (Autshumato)	19.56	Scraping by
English	Northorn Sotho (JW300)	15.40	Scraping by
English	Sesotho	41.23	Scraping by
English	Setswana	19.66	Hopeful
English	Shona	30.84	Scraping by
English	Southern Ndebele (I)	4.01	Left Behind
English	Southern Ndebele (II)	26.61	Left Behind
English	kiSwahili (JW300)	48.94	Rising Star
English	kiSwahili (SAWA)	3.60	Rising Star
English	Tigrigna (JW300)	4.02	Hopeful
English	Tigrigna (JW300+Tatoeba+more)	14.88	Hopeful
English	Tiv	44.70	Left Behind
English	Tshiluba	42.52	Left Behind
English	Tshivenda	49.57	Scraping by
English	Urhobo	28.82	Left Behind
English	isiXhosa (Autshumato)	13.32	Hopeful
English	isiXhosa (JW300)	6.00	Hopeful
English	Xitsonga (JW300)	4.44	Scraping by
English	Xitsonga (Autshumato)	13.54	Scraping by
English	Yoruba	38.62	Rising Star
English	isiZulu (Autshumato)	1.96	Hopeful
English	isiZulu (JW300)	4.87	Hopeful
Efik	English	33.68	Winner
French	Lingala	39.81	Scraping by
French	Swahili Congo	33.73	Left Behind
Hausa	English	25.27	Winner
Yoruba	English	39.44	Winner

Table 4: Benchmarks as of May 28, 2020. If not indicated, training domain is JW300. BLEU scores are computed with Sacrebleu (*tokenize='none'*) on the JW300 test sets. Target languages are categorized according to (Joshi et al., 2020) as of 28 May 2020.