Combating Harmful Hype in Natural Language Processing

Asmelash Teka Hadgu
Lesan AI, DAIR
asme@lesan.ai

Paul Azunre
Ghana NLP
azunre@gmail.com

Timnit Gebru
DAIR
timnit@dair-institute.org

Abstract

In recent years, large multinational corporations have made claims of creating "general purpose" models that can handle many different tasks within natural language processing. Recent works from Meta for example, give the impression that they have nearly solved machine translation tasks for more than 200 languages including 55 African languages. In this paper, we outline the harms speakers of non dominant languages have experienced due to these grandiose and inaccurate claims, ranging from diverting resources from local startups to low quality datasets and models from these corporations. We urge the African NLP and machine learning communities to push back against these claims, and support smaller organizations serving their own communities.

1 Introduction and Related Work

A number of Big Tech companies such as Google and Meta have recently released datasets and models for multilingual machine translation, with the languages they are targeting increasing in number Bapna et al. (2022); Lample et al. (2018); NLLB Team et al. (2022); Radford et al. (2022). For instance, in Meta's 2022 paper, "No Language Left Behind: Scaling Human-Centered Machine Translation," the authors ask: "What does it take to break the 200 language barrier while ensuring safe, high quality results, all while keeping ethical considerations in mind?" In the blogposts accompanying these research papers, the companies give the impression that they have nearly solved the task of machine translation from any source language to any target language, stressing their state-of-the-art performance in neglected languages such as those in Africa. Meta's blogpost accompanying NLLB Team et al. (2022) notes [1]:

> We've built a single AI model called NLLB-200, which translates 200 different languages with state-of-the-art results...Fewer than 25 African languages are currently supported by widely used translation tools — many of which are of poor quality. In contrast, NLLB-200 supports 55 African languages with high-quality results.

Similarly, in "Unlocking Zero-Resource Machine Translation to Support New Languages in Google Translate," the authors note that "the languages that are currently represented are overwhelmingly European, largely overlooking regions of high linguistic diversity, like Africa and the Americas," and discuss key challenges "towards building functioning translation models for the long tail of languages." [2] After introducing their approach for creating monolingual datasets and "zero-resource translation," i.e., translation for languages with no in-language parallel text, they note that "the quality of translations produced by these models still lags far behind that of the higher-resource languages supported by Google Translate." While the blog post headline highlights "zero-resource machine translation" allowing Google translate to support more languages, and the body stresses the importance of supporting the

[1] https://ai.facebook.com/blog/nllb-200-high-quality-machine-translation
[2] https://ai.googleblog.com/2022/05/24-new-languages-google-translate.html

"long tail" of languages, the low quality of the translations by the introduced methodology is relegated to being mentioned in passing in the end.

We argue that this type of presentation of results is harmful to the indigenous African language research ecosystem. It deceptively inflates public perception of the quality of these models and datasets for underrepresented languages. It misleads the public into believing that Big Tech companies like Google and Meta have solved machine translation with one model for every language. This in turn siphons resources away from smaller organizations like Ghana NLP [3] and Lesan AI [4] who create machine translation systems for specific communities they belong to. Machine translation tools built by Big Tech companies targeting speakers of these languages are often provably inferior to solutions built by indigenous communities. Yet, local organizations lose resources due to the false impression created by the aforementioned hype, that there is no longer a need for them. Indeed, a number of scholars in NLP have recently pushed back against this hype. For instance, in Bender & Koller (2020), the authors dedicate a section to "hype and analysis." An increasing number of scholars have also analyzed the quality of datasets used in NLP systems Dodge et al. (2021); Kreutzer et al. (2022). We add to this body of work by dissecting hype vs reality. In the rest of the paper, we specifically analyze the datasets and models introduced by NLLB Team et al. (2022), in 4 African languages: Tigrinya, Amharic, Afaan Oromo and Twi. Our preliminary work finds a number of quality issues, including:

- Top dataset sources from websites not hosted in countries with native speakers.
- Potential inclusion of outputs of machine translation systems in the datasets.
- Dataset sources that are not representative of local contexts.
- Very poor model performance in colloquial settings.

2 Preliminary Analysis

In this section, we inspect the quality of the dataset and models released by the NLLB project for four widely spoken African languages: Amharic (approximately 32 million native speakers), Afaan Oromo (approximately 37 million native speakers), Tigrinya (approximately 7 million native speakers), and Twi (approximately 9 million speakers) Ado et al. (2021); Ethnologue. We then discuss how our results compare to the claims on Meta's blogspot[5]. Namely, their claims that "NLLB-200 supports 55 African languages with high-quality results. For some African and Indian languages, the increase is greater than 70 percent over recent translation systems."

The NLLB project uses data from three sources to train and evaluate their machine translation systems: public bitext, mined bitext and data generated using backtranslation NLLB Team et al. (2022). Details about the different data sources and open source links to access them are provided in the NLLB dataset, and we perform analysis on the mined bitext and the FLORES subsets of the NLLB datasets.

2.1 Training Data Quality

At first glance, we found that many of the Websites that were crawled for Tigrinya, Amharic or Afaan Oromo languages are not hosted in the country or region of the native speakers of these languages: Ethiopia and Eritrea for (Tigrinya), Ethiopia and Kenya for Afaan Oromo, and Ethiopia for Amharic. For instance, many domains resolve to `rayhaber.com` or `martech.zone` corresponding to websites hosted in Turkey and The United States respectively. To do this analysis, we used tldextract, a Python library that separates a URL's subdomain, domain, and public suffix, to extract domains from URLs, and used the Public Suffix List (PSL). With this approach, the 3 URLs listed below, all correspond to the domain `www.am.rayhaber.com`.

[3] `https://ghananlp.org/`

[4] `https://lesan.ai/`

[5] `https://ai.facebook.com/blog/nllb-200-high-quality-machine-translation`

1 https://am.rayhaber.com/2019/09/erciyesab-ስኪ-ሪዞርት-የ-2020- ወቅት-የኬብል-መኪና-ትኬት-ዋጋዎች።

2 https://am.rayhaber.com/2019/10/ፕሬዝዳንት-ኢማሞግ-እንደ-አንበሶች-ያሉ-ታሪካዊ-ሕንፃዎች-ይኖሯቸዋል/

3 https://am.rayhaber.com/2019/09/ፕሬዝዳንት-ኢሞራሉ-ኢብቢዬ-ወደ-ባህር-ዳርሳ-አውቶቡስ-ጣቢያ-ተዛወሩ-፡ ፡/

Since 7,946,698 (49%) of the Amharic sentences in the Amharic-English bitext contain source URLs, we can analyze the top websites used as sources for this dataset. Table 1 shows the top eight domains that contribute 1,577,139 (19.84%) sentences. Since none of these websites have organic (human generated) Amharic text, we infer that what is included in the dataset is the output of machine translation systems such as Google Translate with source sentences coming from these websites. This is problematic because the output text is not representative of how people use the language naturally, making it inappropriate as training data for machine translation systems.

Source Domain Amharic	**No. of Amharic Sentences**
www.am.econologie.com	541,114
www.am.martech.zone	271,533
www.2fish.co	214174
www.am.rayhaber.com	156,168
www.am.inditics.com	128,746
www.am.everydayprayerguide.com	102,211
www.am.eturbonews.com	90,863
www.actualidadiphone.com	72,330

Table 1: Domain name and number of sentences from each domain for the Amharic subset of the Amharic-English bitext data source in the NLLB dataset.

In this part of our data inspection we wanted to understand whether or not there was cross-language contamination: that is, what percentages of the sentences that were classified as being in Amharic, Tigrinya, and Afaan Oromo were in fact in these languages? To answer this question, we used a language identification tool from Hadgu et al. (2021) to classify the languages of the Amharic, Tigrinya and Afaan Oromo subsets in the dataset. We found that 1.2%, 2%, and 3.2% of the sentences in the Amharic, Tigrinya and Afaan Oromo collections respectively are not from the classified language (see Table 2). We manually inspected a random subset of these sentences to verify that they were indeed not from the intended languages.

Amharic		**Afaan Oromo**		**Tigrinya**	
am	15,941,105	om	3,128,055	ti	1,369,736
ti	133,673	en	100,691	am	20,310
und	61,515	und	3,705	und	8,053
om	582	am	46	en	44
en	178	ti	16	om	30

Table 2: The output of the Hadgu et al. (2021) language identification tool when applied to the subsets of the NLLB bitext dataset tagged as Amharic in the Amharic-English pair (1^{st} column), Afaan Oromo in the English Afan Oromo pair (2^{nd} column) and Tigrinya in the English-Tigrinya pair (3^{rd} column). am=Amharic, ti=Tigrinya, und=Unknown, om=Afaan Oromo, en=English

2.2 Evaluation Data Quality

The FLORES-200 is an evaluation benchmark for low-resource and multilingual machine translation from Meta. This is the main evaluation dataset used by NLLB to report results.

Source Amharic	English Translation
የሳይንስ ሊቃውንት በግጭቱ ምክንያት የተፈጠረው ፍንዳታ ከፍተኛ እንደነበር ይናገራሉ።	Scientists say the explosion caused by the collision was massive.
ፖሊስ ባለው መሰረት፣ ፎቶግራፈሩን የገጨው ተሽከርካሪ ሹፌር የወንጀለኛ ቅጣት የመቀበል ዕድሉ እጅግ አናሳ ነው።	According to police, the driver of the vehicle that hit the photographer is unlikely to face criminal charges.
ታዋቂ ስፖርቶች እግር ኳስ፣ ቅርጫት ኳስ፣ መረብ ኳስ፣ ውሃ-ፖሎ፣ የጎራዴ ጨዋታ፣ ራግቢ፣ ሳይክል መንዳት፣ የበረዶ ሆኪ፣ ሮለር ሆኪ እና ኤፍ1 የሞተር ውድድርን ያካትታሉ።	Popular sports include football, basketball, volleyball, water-polo, fencing, rugby, cycling, ice hockey, roller hockey and F1 motor racing.
ሁሉም ሰው በኅብረተሰብ ውስጥ ይሳተፋል እና የትራንስፖርት ስርዓቶችን ይጠቀማል። ሁሉም ሰው ማለት ይቻላል ስለ መጓጓዣ ስርዓቶች ቅሬታ ያቀርባል።	Everyone participates in society and uses transportation systems. Almost everyone complains about transportation systems.
አንድ ድርጅት ፈጠሪ ከመሆኑ በፊት አመራር የፈጠራ ስራ ባህል እንዲሁም የጋራ ዕውቀት እና ድርጅታዊ ትምህርት መፍጠር አለበት።	Before an organization can be innovative, leadership must create a culture of innovation as well as shared knowledge and organizational learning.

Table 3: Examples form FLORES-200 dataset where translation matches 100% with the output of Google Translate.

Location (LOC)	Organization (ORG)	Person (PER)
Trafalgar Square	Harvard Law School	Arthur Guinness
The Oyapock River Bridge	Turkish Airlines	Jonny Reid
Galapagos	the Minneapolis Star-Tribune	Brzezinski
		John F. Kennedy

Table 4: Sample named entities in the FLORES 200 dataset.

While running machine translation startups serving African languages, we have seen annotators we hire to generate ground truth first use Google Translate from the source language (e.g. English) to the target language (e.g. Amharic, Tigrinya, Afaan Oromo), and then edit the output to generate the final sentences. This type of data leak could skew evaluation systems because instead of generating ground truth data from scratch, the evaluation data would also be generated by the model that is supposed to be evaluated (such as Google Translate). In this section, we seek to understand whether or not a subset of the data is the output of a machine translation system rather than human generated text.

While we are still working on quantifying the percentage of text in the dataset that is the output of another automated translation system, we have observed a number of Amharic-English sentence pairs, where the Amharic text seems to be taken from a Google Translate output. Table 3 shows examples of Amharic-English pairs where the English translation matches 100% with the output of Google Translate when the source sentences are the ones shown in Table 3. While we have no way of saying for certain whether the target sentences are outputs from Google Translate, it is highly unlikely that annotators would have exactly the same output word for word, which leads us to suspect that this is indeed the case.

An evaluation dataset for African languages should ensure that local contexts and named entities are appropriately translated. That is, when translating from a language like Tigrinya to English (for example), we would expect the local context to contain named entities such as places of interest or prominent individuals, who should be recognized as such in the English translation. If the named entities in English that one includes in their dataset is based on Wikipedia entries, like FLORES-200 does, however, those considered to be prominent will only be names and individuals that are recognized as such in Wikipedia articles, which have a known Western and male bias Barera (2020).

Table 4 shows a random sample of named entities in the FLORES dataset, with almost no coverage of entities that would be locally relevant for languages such as Amharic, Afaan Oromo or Tigrinya. This makes the dataset less relevant for evaluating machine translation systems whose source texts are in these three languages, and consist of local named entities. We are working on quantifying this disparity in the named entity representation.

2.3 Model Quality

While we performed a preliminary analysis of subsets of the NLLB dataset for Amharic, Tigrinya and Afaan Oromo in the prior sections, here we evaluate the quality of the machine translation systems in the NLLB project for Twi.

NLLB has two relevant categories for Twi: aka_Latn (ak) which is supposed to represent the Asante dialect Paster (2010) and twi_Latn (tw) which is supposed to represent the Akuapem dialect Tuffour (2020). These are referred to as "Akan" and "Twi" respectively in the NLLB paper and documentation (also evidenced by the language codes chosen - ak and tw respectively). The chosen language codes, ak and tw, and their descriptions as Akan and Twi, show a basic cultural misunderstanding about these languages: both of these dialects are languages of the Akan people, and both of them are also Twi. This misunderstanding of a basic cultural and linguistic fact, built directly into the model, raises serious questions.

We evaluated the performance of the NLLB-200-distilled-600M model on two benchmarks: 1) The crowdsourced subset of the Twi dataset from Azunre et al. (2021) (mostly consists of Asante Twi), and 2) The test split of the Twi component of Adelani et al. (2022). The 600M variant of the NLLB model was chosen because it has a comparable number of parameters to the distilled Google Translate models Bapna et al. (2022). Note that the first benchmark dataset is mostly colloquial everyday speech while the second one consists of news stories. We compare performance to models from Khaya and Google Translate for both directions. The results are listed below.

Model	Colloquial BLEU	News BLEU
Khaya(en-tw)	15.2	8.47
Google (en-tw)	15.6	8.62
NLLB "ak" (en-tw)	1.71	6.10
NLLB "tw" (en-tw)	8.12	7.02
Khaya (tw-en)	29.7	8.30
Google (tw-en)	28.0	18.4
NLLB "ak" (tw-en)	6.90	10.3
NLLB "tw" (tw-en)	2.95	8.38

Table 5: Comparison of NLLB model performance with Google Translate and Khaya.

While the NLLB models achieve a BLEU score of more than 7 on the news corpus, on colloquial speech—which is where many people apply machine translation systems in their daily lives–they achieve scores as low as 1.71.

3 Conclusion and Future work

Looking at dataset quality in Afaan Oromo, Tigrinya and Amharic, it is clear that Meta did not perform extensive and thorough analysis before making claims of high quality results. Similarly, analyzing model performance on multiple Twi dialects, it is clear that Meta did not test these models widely enough on relevant benchmarks before making claims of high quality results. These types of grandiose claims result in material harms both to speakers of these languages –who are served by subpar products that are advertised as high quality, and organizations that specifically serve these groups. One potential outcome is reduced investment and support for these organizations due to inaccurate claims that the problem of machine translation is solved. Such claims can negate the need for local startups in the minds of various stakeholders. We urge the machine learning community, and the African NLP community in particular, to push back against such harmful hype.

References

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. 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*, pp. 3053–3070, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.223. URL https://aclanthology.org/2022.naacl-main.223.

Derib Ado, Almaz Wasse Gelagay, and Janne Bondi Johannessen. The languages of ethiopia. *Grammatical and Sociolinguistic Aspects of Ethiopian Languages*, 48:1, 2021.

Paul Azunre, Lawrence Adu-Gyamfi, Esther Appiah, Felix Akwerh, Salomey Osei, Cynthia Amoaba, Salomey Afua Addo, Edwin Buabeng-Munkoh, Nana Boateng, Franklin Adjei, and Bernard Adabankah. English-akuapem twi parallel corpus, January 2021. URL https://doi.org/10.5281/zenodo.4432117.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*, 2022.

Michael Barera. Mind the gap: Addressing structural equity and inclusion on wikipedia. 2020.

Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5185–5198, 2020.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus, 2021. URL https://arxiv.org/abs/2104.08758.

Ethnologue. Languages of ghana.

Asmelash Teka Hadgu, Abel Aregawi, and Adam Beaudoin. Lesan - machine translation for low resource languages. In *NeurIPS 2021 Competitions and Demonstrations Track, 6-14 December 2021, Online*, volume 176 of *Proceedings of Machine Learning Research*, pp. 297–301. PMLR, 2021. URL https://proceedings.mlr.press/v176/hadgu22a.html.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation, 2018. URL https://arxiv.org/abs/1804.07755.

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. No language left behind: Scaling human-centered machine translation. 2022.

Mary Paster. The verbal morphology and phonology of asante twi. *Studies in African Linguistics*, 39(1):78–99, 2010.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL `https://arxiv.org/abs/2212.04356`.

Adu David Tuffour. Comparative and contrastive analysis of vowel harmony in asante and akuapem twi dialects in ghana. *International Journal of Research and Scholarly Communication*, 3(1):42–51, 2020.