INMT-*Lite*: Accelerating Low-Resource Language Data Collection via Offline Interactive Neural Machine Translation

LREC-COLING 2024

Harshita Diddee[*♡2], Anurag Shukla[*1], Tanuja Ganu[1], Vivek Seshadri[1]
Sandipan Dandapat[3], Monojit Choudhury[4♡], Kalika Bali[1]
Carnegie Mellon University[2], Microsoft Research India[1], Microsoft R&D, India[3]
Mohamed Bin Zayed University Of Artificial Intelligence [4]
hdiddee@andrew.cmu.edu, monojit.choudhury@mbzuai.ac.ae
{anuragshukla, visesha,taganu, sadandap,kalikab}@microsoft.com

Context

1. Advanced Multilingual models does not benefit all languages equally.[1][2]
2. They especially fail for low resource languages.
3. It is often due to lack of data available for these languages. [3][4]

Data Collection

Often done in two ways:

1. Without Assistance
2. With Assistance
    a. Interactive
    b. Non Interactive

INMT (Interactive Machine Translation)

It is a form of assistive translation where the probability of next token does not depend on output tokens of previous invocation but instead depend on tokens being generated by the user.

Cont…

1. Aims to provide “Interactive” assistance during translation
2. Improves yield of data collected
3. Enhances the overall experience of data providers

Problems with INMT

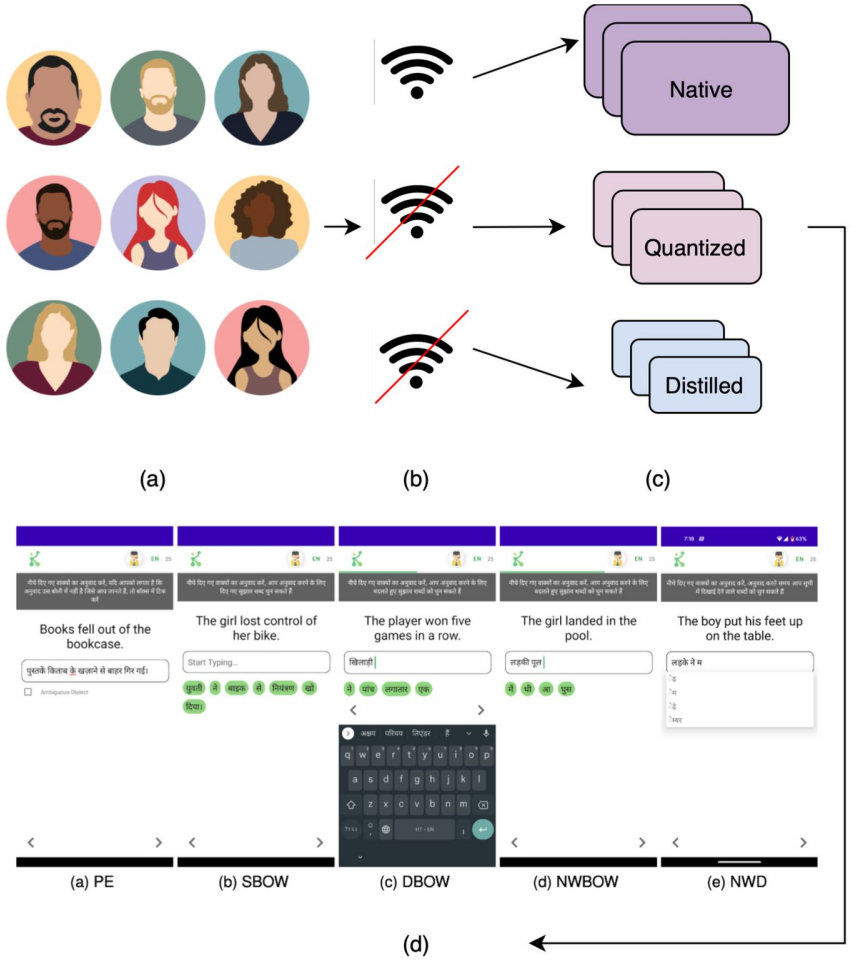
Under resourced language communities often does not have access to internet and high-end devices.

“INMT often ignores the environment where such system can be used to increase the yield of data”

INMT-lite

INMT-lite:

1. Adapts to the infrastructural capabilities of the user.
    a. Deployed on edge
    b. Works without internet
    c. Works on a smartphone
2. Provides several user interfaces to account for the quality of underlying model.


Native
Quantized
Distilled
(a)
(b)
(c)
Books fell out of the bookcase.
The girl lost control of her bike.
The player won five games in a row.
The girl landed in the pool.
The boy put his feet up on the table.
(a) PE
(b) SBOW
(c) DBOW
(d) NWBOW
(e) NWD
(d)

Models

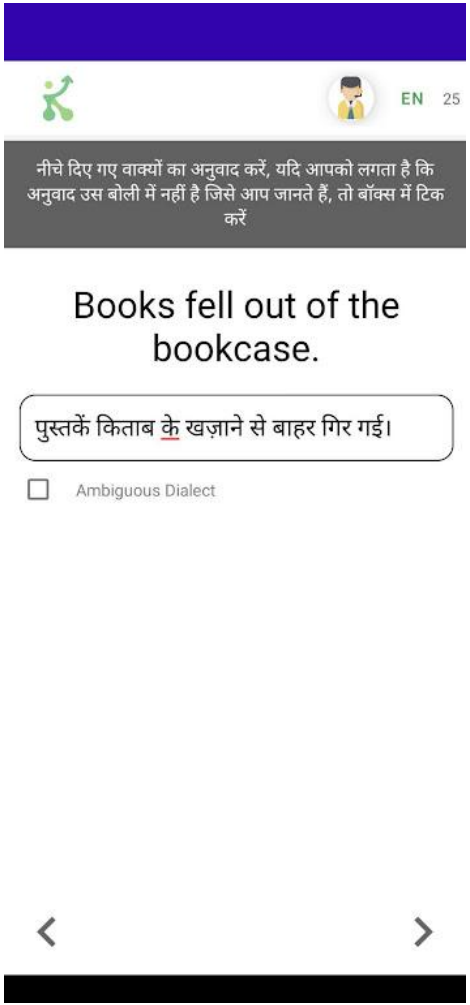
1. Native (Internet-enabled, uncompressed)
2. Quantized (Internet-Independent, Compressed Model)
3. Distilled (Internet-Independent, Compressed Model)

Language	Data	M		Q(M)	
		BLEU	*chrF*	*BLEU*	*chrF*
Punjabi	2.4M	38.4	50.6	27.0	48.0
Gujarati	3.0M	35.9	53.4	28.4	51.4
Marathi	3.3M	27.5	52.7	11.1	40.8
Bengali	8.4M	24.9	46.8	11.4	35.1
Hindi	8.5M	37.7	59.9	27.1	44.9

Language	Data	M		D(M)	
		BLEU	*chrF*	*BLEU*	*chrF*
Gondi	25K	14.3	32.5	14.2	32.8
Assamesse	140K	10.4	30.4	9.6	27.4
Odia	1M	27.4	47.6	20.2	40.7

Interfaces

POST EDIT


EN 25
नीचे दिए गए वाक्यों का अनुवाद करें, यदि आपको लगता है कि अनुवाद उस बोली में नहीं है जिसे आप जानते हैं, तो बॉक्स में टिक करें
Books fell out of the bookcase.
पुस्तकें किताब के खज़ाने से बाहर गिर गई।
Ambiguous Dialect

STATIC BOW

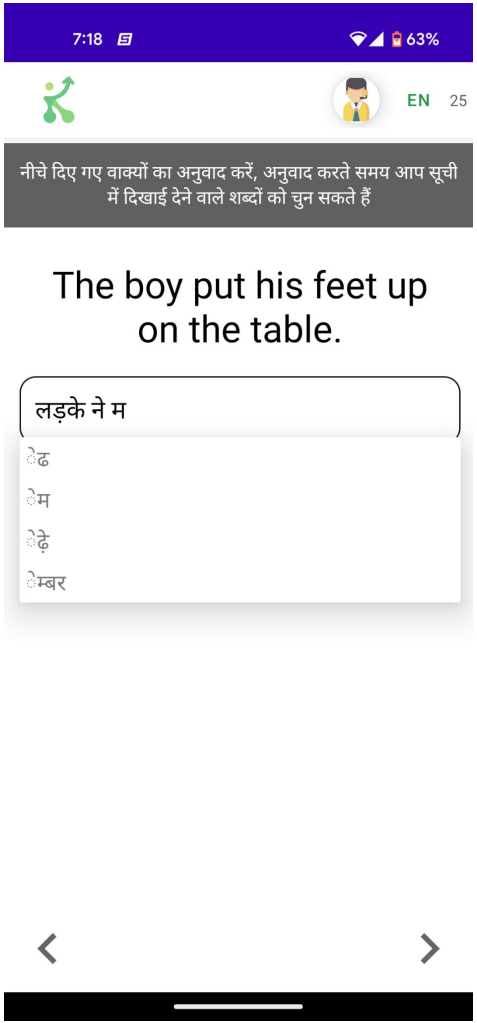

EN 25
नीचे दिए गए वाक्यों का अनुवाद करें, आप अनुवाद करने के लिए दिए गए सुझाव शब्द चुन सकते हैं
The girl lost control of her bike.
Start Typing...
युवती
ने
बाइक
से
नियंत्रण
खो
दिया।

DYNAMIC BOW


EN 25
नीचे दिए गए वाक्यों का अनुवाद करें, आप अनुवाद करने के लिए बदलते हुए सुझाव शब्दों को चुन सकते हैं
The player won five games in a row.
खिलाड़ी
ने
पांच
लगातार
एक
अक्षय
परिचय
लिएंडर
हैं
HT • EN

NEXT WORD BOW

NEXT WORD DROPDOWN



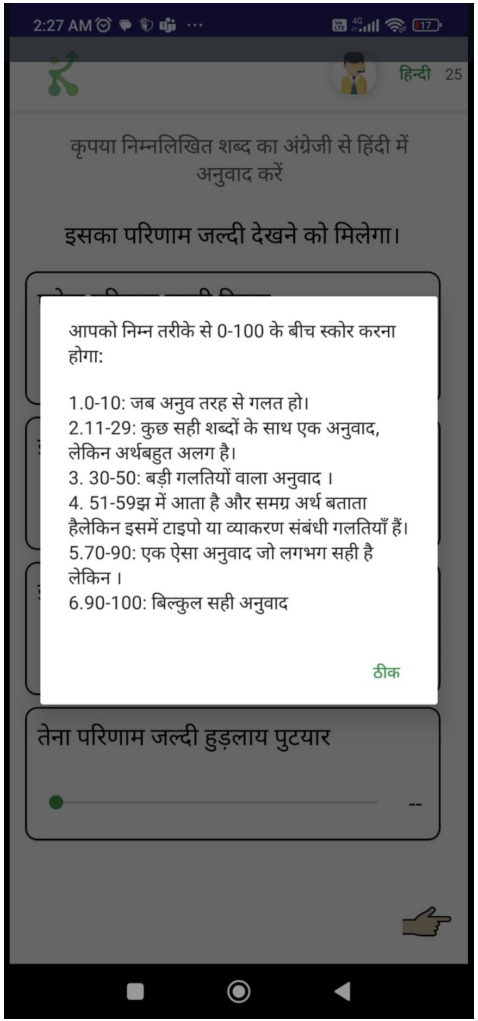
User Study

Conducted an extensive user study with the Gond community in Chattisgarh province of India. We went through following high level steps:

1. Data collected from 18 annotators for all the interfaces
2. Scoring of the translation by direct assessment (DA) scores
3. Feedback from the annotators

Task Setup

Task Name	Interface	Task Description
Baseline	Default	Users provide translations without any assistance.
Assistive	Default, Bag of Words and Dropdown	Users provide translations by post-editing or using the model's assistance via each of the assistive interfaces.
Scoring	DA Scoring	Users score the translations generated by users. The highest ranked translation here will then further be marked to identify the best mode.


2:27 AM
हिन्दी 25
कृपया निम्नलिखित शब्द का अंग्रेजी से हिंदी में अनुवाद करें
इसका परिणाम जल्दी देखने को मिलेगा।
आपको निम्न तरीके से 0-100 के बीच स्कोर करना होगा:
1.0-10: जब अनुव तरह से गलत हो।
2.11-29: कुछ सही शब्दों के साथ एक अनुवाद, लेकिन अर्थबहुत अलग है।
3. 30-50: बड़ी गलतियों वाला अनुवाद ।
4. 51-59झ में आता है और समग्र अर्थ बताता हैलेकिन इसमें टाइपो या व्याकरण संबंधी गलतियाँ हैं।
5.70-90: एक ऐसा अनुवाद जो लगभग सही है लेकिन ।
6.90-100: बिल्कुल सही अनुवाद
ठीक
तेना परिणाम जल्दी हुड़लाय पुटयार
--

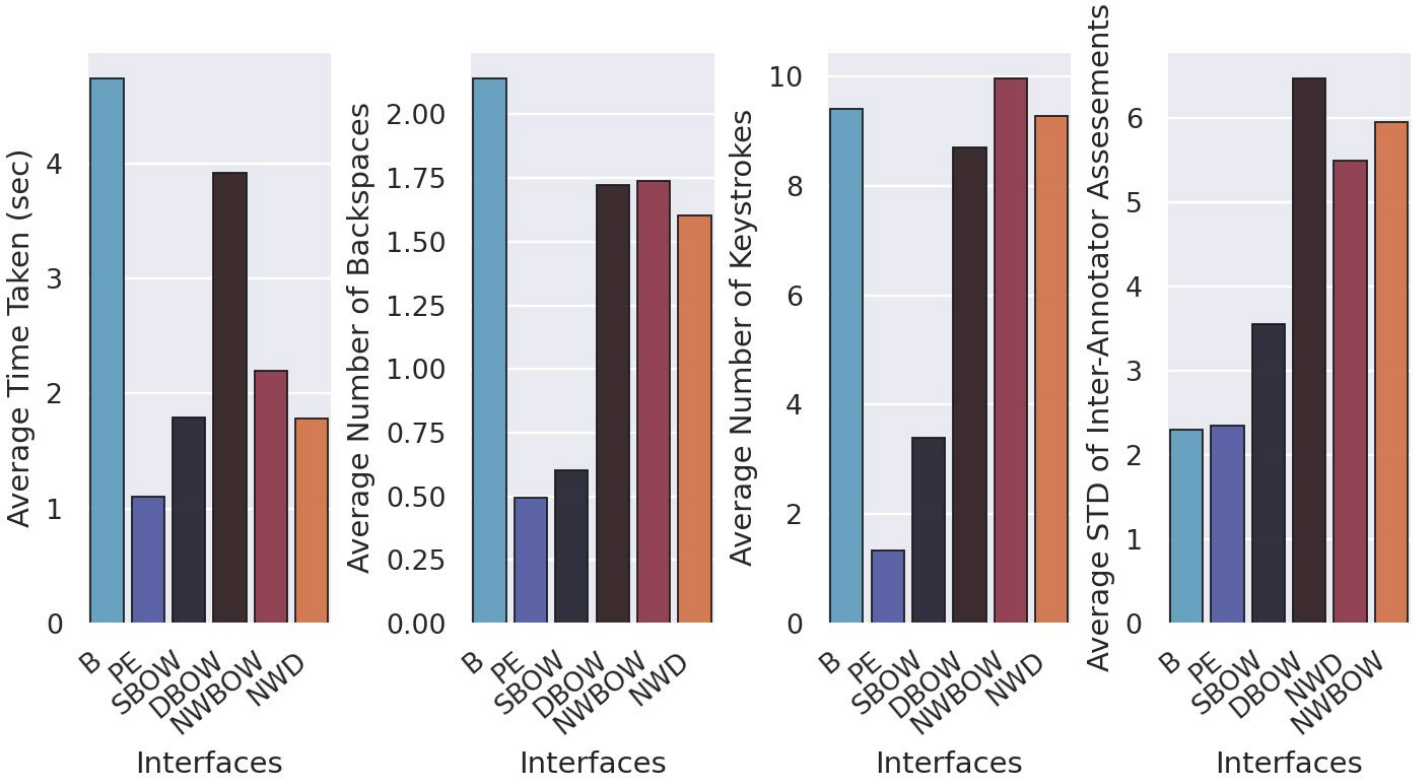
Distribution Strategy

Annotator × Task	a_1	a_2	a_3	a_4	a_5	a_6
t_1	(s_1, i_1)	(s_1, i_2)	(s_1, i_3)	(s_1, i_4)	(s_1, i_5)	(s_1, i_6)
t_2	(s_2, i_2)	(s_2, i_3)	(s_2, i_4)	(s_2, i_5)	(s_2, i_6)	(s_2, i_1)
t_3	(s_3, i_3)	(s_3, i_4)	(s_3, i_5)	(s_3, i_6)	(s_3, i_1)	(s_3, i_2)
t_4	(s_4, i_4)	(s_4, i_5)	(s_4, i_6)	(s_4, i_1)	(s_4, i_2)	(s_4, i_3)
t_5	(s_5, i_5)	(s_5, i_6)	(s_5, i_1)	(s_5, i_2)	(s_5, i_3)	(s_5, i_4)
t_6	(s_6, i_6)	(s_6, i_1)	(s_6, i_2)	(s_6, i_3)	(s_6, i_4)	(s_6, i_5)

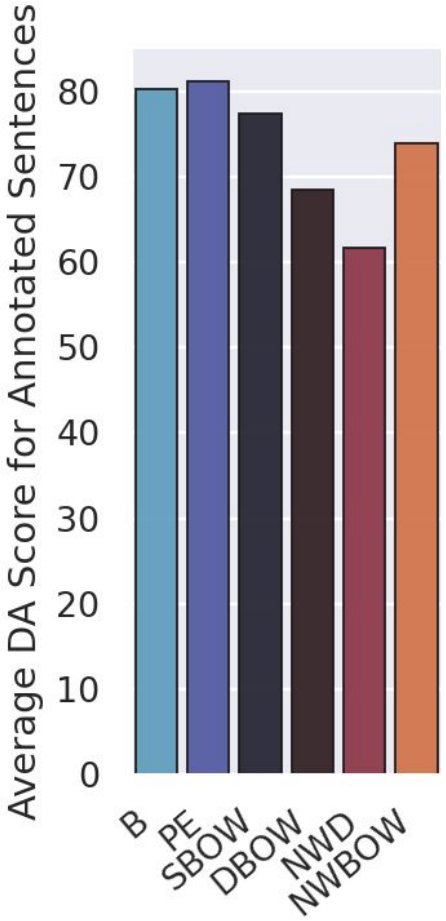
Questions we are trying to answer

1. Does INMT-lite lead to reduction in human effort?
2. How well do the translations generated by INMT-lite compare with those generated without assistance?
3. Does INMT Lite improve the experience of annotators during data generation

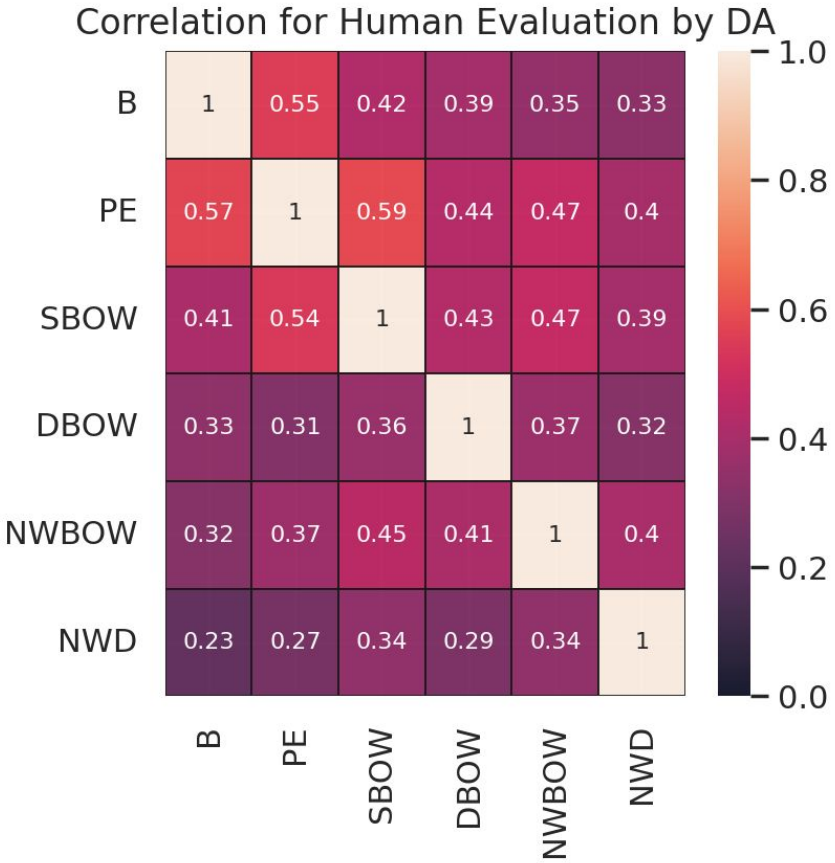
Does INMT-lite lead to reduction in human effort?

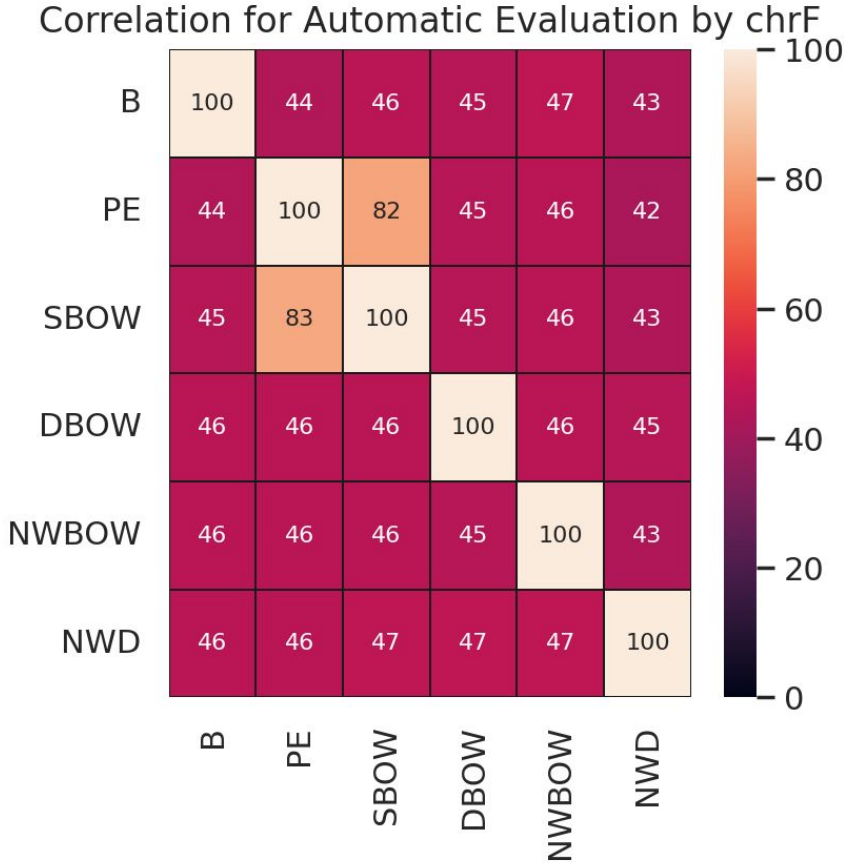


Assistance vs No Assistance



Cont…


Correlation for Human Evaluation by DA
B 1 0.55 0.42 0.39 0.35 0.33
PE 0.57 1 0.59 0.44 0.47 0.4
SBOW 0.41 0.54 1 0.43 0.47 0.39
DBOW 0.33 0.31 0.36 1 0.37 0.32
NWBOW 0.32 0.37 0.45 0.41 1 0.4
NWD 0.23 0.27 0.34 0.29 0.34 1
B PE SBOW DBOW NWBOW NWD
1.0 0.8 0.6 0.4 0.2 0.0


Correlation for Automatic Evaluation by chrF
B 100 44 46 45 47 43
PE 44 100 82 45 46 42
SBOW 45 83 100 45 46 43
DBOW 46 46 46 100 46 45
NWBOW 46 46 46 45 100 43
NWD 46 46 47 47 47 100
B PE SBOW DBOW NWBOW NWD
100 80 60 40 20 0

Experience by Annotators

1. Breadth-wise coverage interfaces (SBOW, DBOW and PE) more helpful than depth-wise coverage interfaces (NBOW).
2. Annotators often prefered typing the suggestion instead of selecting it.
3. The suggestions often helped them jump start their translation.

Future Work

INMT-lite operates on vast set of parameters that needs to be investigated further:

1. Depth of decoding.
2. The number of suggestions shown across each architecture.
3. The trigger of invocation.

Thank You!

https://github.com/microsoft/INMT-lite

References

1. Daniel J. Liebling, Michal Lahav, Abigail Evans, Aaron Donsbach, Jess Holbrook, Boris Smus, and Lindsey Boran. 2020. Unmet needs and opportunities for mobile translation ai. In Pro- ceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI ’20, page 1–13, New York, NY, USA. Association for Com- puting Machinery.
2. 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 Onyefu- luchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Ök- tem, 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.
3. Devansh Mehta, Harshita Diddee, Ananya Sax- ena, Anurag Shukla, Sebastin Santy, Ramar- avind Kommiya Mothilal, Brij Mohan Lal Srivas- tava, Alok Sharma, Vishnu Prasad, U. Venkanna, and Kalika Bali. 2022. Learnings from techno- logical interventions in a low resource language: Enhancing information access in gondi. ArXiv, abs/2211.16172.
4. David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for so- cially equitable language identification. In Pro- ceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol- ume 2: Short Papers), pages 51–57, Vancouver, Canada. Association for Computational Linguis- tics.