Machine Translation Of Marathi Dialects: A Case Study Of Kadodi

Raj Dabre[1,2] **Mary Noel Dabre**[3] **Teresa Pereira**[4]
National Institute of Information and Communications Technology, Kyoto, Japan[1]
IIT Madras, Chennai, India[2]
Independent, Vasai, India[3]
St Gonsalo Garcia College, Vasai, India[4]
raj.dabre@nict.go.jp

Abstract

While Marathi is considered as a low- to middle-resource language, its 42 dialects have mostly been ignored, mainly because these dialects are mostly spoken and rarely written, making them extremely low-resource. In this paper we explore the machine translation (MT) of Kadodi, also known as Samvedi, which is a dialect of Marathi. We first discuss the Kadodi dialect, highlighting the differences from the standard dialect, followed by presenting a manually curated dataset called ***Suman*** consisting of a trilingual Kadodi-Marathi-English dictionary of 949 entries and 942 simple sentence triples and idioms created by native Kadodi speakers. We then evaluate 3 existing large language models (LLMs) supporting Marathi, namely Gemma-2-9b, Sarvam-2b-0.5 and LLaMa-3.1-8b, in few-shot prompting style to determine their efficacy for translation involving Kadodi. We observe that these models exhibit rather lackluster performance in handling Kadodi even for simple sentences, indicating a dire situation.

1 Introduction

Marathi is a language primarily spoken by about 83 million people[1] in the Indian state of Maharashtra. Across the world, while a standard dialect of any language exists, a substantial portion of these speakers also speak a local dialect and Marathi is no exception. There are 42 known dialects of Marathi[2] a vast majority of which, if not all, are spoken rather than written, which makes natural language processing (NLP) for such dialects extremely hard. However, excluding these dialects from NLP systems would lead to a cultural representation imbalance, since a significant amount of culture is connected to languages and their dialects. Given the massive Marathi dialect-speaking population, we consider it important to take steps to include them in NLP systems, the first being via resource creation and evaluation.

In this paper we focus on a minor dialect of Marathi, namely, Kadodi[3], also known as Samvedi, which is spoken in the Vasai[4] region of Maharashtra and has about 60,000 native speakers. The Kadodi language is a mix of Konkani, Gujarati, Marathi and Indo-Portuguese (now extinct). The speakers of Kadodi are known colloquially as Kuparis[5][6] which essentially means comrade and is a term used to call one's child's godfather. The Kupari people are descendants of a mixture of Samvedi Brahmins, Goan Konkani Brahmins and Portuguese New Christians; because of intermarriages between them. Due to it being a spoken dialect, it has been passed down over the generations mainly via conversations. However, this also means that there is no proper text data available for NLP applications.

In this paper, we present the first of its kind study of Kadodi taking Machine Translation (MT) as a NLP application. We first describe the features of the Kadodi language and explain its differences from Marathi. Then, we describe the process of data collection, which was mainly done via two native speakers of Kadodi, leading to ***Suman***, the first tri-parallel Kadodi-Marathi-English dataset. Finally, we attempt to evaluate the translation quality of Kadodi translation both to and from English and Marathi via few-shot prompting of 3 LLMs. where we show that despite our evaluation being conducted on simple sentences, all LLMs we considered exhibit lackluster performance, indicating the need for dedicated pre-training and fine-tuning

[1] https://en.wikipedia.org/wiki/Marathi_language

[2] https://en.wikipedia.org/wiki/Marathi_language#Dialects

[3] https://en.wikipedia.org/wiki/Kadodi_language

[4] https://en.wikipedia.org/wiki/Vasai

[5] https://en.wikipedia.org/wiki/Kupari

[6] The feminine form of Kupari is Kumari.

Proceedings of the Eleventh Workshop on Asian Translation (WAT 2024), pages 36–44

on dialectic data. Our contributions are as follows:

1. The first study of Kadodi machine translation.

2. A novel dataset called ***Suman***, for Kadodi-Marathi-English 3-way parallel entries with about 1,900 dictionary and sentence pairs, totally. We release our dataset publicly[7].

3. An evaluation of the translation quality of existing models involving Kadodi.

Going forward, Kadodi refers to the Kadodi dialect and Marathi refers to the standard dialect.

2 Related Work

This paper mainly focuses on the natural language processing of dialects, specifically machine translation involving the Kadodi dialect of Marathi.

A vast majority of the dialectic work has been conducted on Arabic, English and French dialects, and some of the most prominent works have been on dialect understanding (Baimukan et al., 2022; Zampieri et al., 2014; Malmasi et al., 2016; Goutte et al., 2016; Elmadany et al., 2018; Joukhadar et al., 2019) and dialect translation (Zbib et al., 2012; Bouamor et al., 2018; Contarino, 2021; Lent et al., 2024; Robinson et al., 2024)[8]. On the other hand, works on summarization (Olabisi et al., 2022; Keswani and Celis, 2021) and dialogue (Elmadany et al., 2018; Joukhadar et al., 2019; Marietto et al., 2013) are rather limited due to the unavailability of data or lack of permissive licenses.

Since dialects are closely related to their standard variant, multilingual transfer learning (Dabre et al., 2020) approaches are often helpful alongside approaches leveraging transliteration (J et al., 2024; Dabre et al., 2022). Additionally, character level systems (Abe et al., 2018) are often effective in settings where the training data for dialects is rather limited, where regularization approaches are also effective (Liu et al., 2022; Maurya et al., 2023). In low-resource settings, it becomes important to leverage linguistic features, ideally of dialects, to improve translation quality (Erdmann et al., 2017; Chakrabarty et al., 2022, 2020). On the other hand, since many dialects are spoken, some researcher focus directly on creating and leveraging speech data (Plüss et al., 2023). Joshi et al. (2024) give a comprehensive survey of NLP for dialects across the world, and we encourage readers to read it for an in-depth understanding of the prominent works carried out in this area.

Works on dialects of Indian languages are rather nonexistent, with a few exceptions (Maurya et al., 2023). To the best of our knowledge, this is the first work on machine translation involving Kadodi and in general on any dialect of Marathi.

Kadodi	Marathi	English
लात (lat)	लाथ (lath)	kick
दुद (dud)	दूध (dudh)	milk
ऑजा (auja)	ओझे (ooje)	burden
शार (shaar)	चार (char)	four
हॅन (haen)	शेण (shen)	cowdung
हन (hun)	सण (sun)	festival

Table 1: Representative Kadodi words with their Marathi and English equivalents and pronunciations.

3 *Suman*: A Kadodi Parallel Corpus

We first give details about the Kadodi dialect contrasting it with Marathi followed by a description of the Kadodi parallel corpus we created from scratch, which we refer to as ***Suman***. This consists of a trilingual Kadodi-Marathi-English dictionary and simple, short sentences.

3.1 Kadodi Language

Given that Kadodi is a dialect of Marathi, it exhibits an extremely high degree of similarity with the latter, with very few lexical and grammatical differences. We now briefly explain some key differences as follows:

Vowels and Consonants: Marathi primarily uses 14 vowels[9] and 34 consonants. However, since Kadodi is primarily a spoken language, it does not use 2 out of 14 vowels, namely, ऐ (ay) and औ (au), and 4 out of 34 consonants, namely, च (cha), छ (ccha), ण (na), and ष (sha). The reasons for this is unknown and undocumented due to the spoken nature of Kadodi, but consonant dropping[10] is a common feature in dialects.

[7] https://github.com/prajdabre/kadodinlp

[8] To be accurate, Lent et al. (2024) and Robinson et al. (2024) focus on Creoles and not dialects. However, we list these works as applicable to dialects because of the high similarity between Creoles and their ancestor languages, which is analogous to the similarity between dialects.

[9] Since not everyone is familiar with IPA, we refer readers to take a look here for an easier reference on how to better read these characters.

[10] https://en.wikipedia.org/wiki/Phonological_history_of_English_consonant_clusters

Language	Sentence
Kadodi	तॉ मजुरी करॉन पॉट भरतॅ tou majuri karon pout bhartae
Marathi	तो मजुरी करून पोट भरतो to majuri karun pot bharto
English	He makes a living by working as a laborer
Kadodi	तॉ निजलॅ tou nejlay
Marathi	तो झोपला आहे to zhopla ahe
English	He is sleeping

Table 2: Examples of Kadodi sentences along with their Marathi and English translations and transliterations.

Kadodi Vocabulary: Table 1 gives a list of some words in Kadodi with their pronunciations, alongside Marathi and English translations. The reader will be able to note that the words look mostly similar, and the key differences lies in the consonant usage. For example, the word for cow dung is हॅन (haen) [Marathi word is शेण (shen)], where the key difference is the use of हॅ (hae) in place of शे (she). Note that it is fairly common for श (sha) and स (sa) to be replaced with ह (ha) in Kadodi. Kadodi also differs from Marathi in that it prefers to use voiced or voiceless dental plosives [त (ta) द (da)] instead of aspirated and murmured ones [थ (tha) ध (dha)]. Note that a stark change in consonants does not occur, and often the changes are rather minor. For example, a plan nasal labial consonant will never be replaced by a fricative glottal one.

Kadodi Grammar: In Table 2 we give examples of Kadodi sentences to highlight the subtle differences with Marathi. As can be seen, the Marathi and Kadodi sentences sound mostly similar. The main difference is in the word forms भरतॅ (bharte) vs भरतो (bharto), and the word choices, निजलॅ (nijley) vs झोपला[11] (zhopla). Another interesting difference is that in Marathi we use झोपला आहे (zhopla ahe) to say "(he/she) is sleeping" (present tense) where आहे (ahe) is the verb for "is" or "to be", however, in Kadodi, although आहे (ahe) can be translated as हाय (hai), it is often omitted for the present tense.

Although there are other minor differences between Marathi and Kadodi, we refer the readers to Russell and Cohn (2012); Francis Correia (1992) for detailed overviews. We also point to a book on the Kadodi (Samvedi) community by Pereira (2007). There are also magazines[12] in Kadodi for interested readers.

[11] झोपणे (zhopne) is the more commonly used word for sleeping, whereas निजणे (nizne) is less commonly used in Marathi.

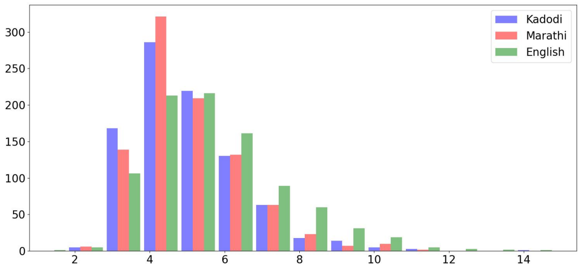

Figure 1: Distribution of Kadodi, Marathi and English sentence lengths.

3.2 Data Collection

We now describe how we collected data for Kadodi MT to create ***Suman***. We primarily focused on collecting Kadodi-Marathi data, since the native speakers (annotators) of both dialects do not possess native English proficiency. The annotators were asked to freely construct any sentences which came to mind, as long as they considered them to be useful in daily conversations. Therefore, the domain of the dataset can be said to be a mix of general domain, conversational and daily use. As much as possible, we asked the annotators to provide English translations, which were manually corrected by native speakers. Annotators were asked to provide dictionary entries as well simple phrases/sentences, leaving longer, complex sentences for the future. All the data was collected over the span of 1 month via Google sheets. We had 2 annotators, and they provided a total of 949 tri-parallel dictionary entries and 942 tri-parallel short sentences. Due to lack of funds, both annotators agreed to create data for free, and for compensation, they were given authorship of this paper.

Dictionary: With the help of annotators, we have procured a dictionary of 949 entries, starting with all 30 consonants and 12 vowels used in Kadodi. Furthermore, the annotators have ensured that for each consonant and vowel type, there are at least 4 Kadodi words. This dictionary also contains roughly 200 instances of numbers, common foods, animals and birds, days of the week, names of months, family relationships, daily use words,

[12] https://kadodi.in/

Shots	kad-mar			kad-eng			mar-kad			eng-kad		
	S	G	L	S	G	L	S	G	L	S	G	L
1	17.0	30.3	37.0	24.3	25.7	28.7	20.2	28.5	30.1	13.2	15.7	18.6
4	22.8	35.4	42.0	24.9	31.4	32.2	18.3	30.4	**33.5**	14.3	15.3	19.4
8	24.4	35.9	42.1	24.3	31.3	32.0	20.0	30.2	32.3	17.1	13.0	**19.6**
12	24.3	36.6	**42.8**	23.1	**33.1**	32.6	18.5	30.2	32.3	16.5	14.2	18.7

Table 3: chrF scores of translation for Kadodi-Marathi (kad-mar), Kadodi-English (kad-eng), Marathi-Kadodi (mar-kad) and English-Kadodi (eng-kad) with 1, 4, 8 and 12 shots. We have compared Sarvam-2b-0.5 (S), Gemma-2-9b (G) and LLaMa-3.1-8b (L) models.

parts of the body, seasons and comparative words. **Sentences:** In addition to the dictionaries, the annotators also created 912 Kadodi sentences of 2199 unique words along with their Marathi and English translations of 1924 and 1650 unique words, respectively. The sentence length distribution is shown in Figure 1. As is evident, most of these are short phrases and sentences between 2 and 6 words, and the length distributions are mostly similar. Note that, Kadodi and Marathi are both morphologically rich languages, so a word can often be the equivalent of a sentence via agglutination. Therefore, just because the sentence lengths appear to be short, they are not all necessarily short in the content they encapsulate. The annotators also created 30 Kadodi idioms along with their literal Marathi translations and explanations in Marathi and English, leading to 942 triples. However, we do not consider these for our experiments.

4 Experiments

We now describe some simple experiments we conduct for Kadodi⇔English and Kadodi⇔Marathi translation using LLMs.

4.1 Settings

For our experiments, we only focus on the parallel sentences part of ***Suman***. Of the 942 Kadodi-Marathi-English triples, we randomly choose 12 triples for 1, 4, 8 and 12-shot prompting and set them aside. Note, once again, we also set aside 30 idiom triples. This leaves us with 900 triples for testing. As for the models, we use Sarvam-2b-v0.5[13] a 2 billion parameter model, Gemma-2-9b (Team et al., 2024) a 9 billion parameter model, and LLaMA-3.1-8b (Dubey et al., 2024) an 8 billion parameter model. All 3 models have seen Indian languages during pre-training although, Sarvam-2b-0.5 has been trained exclusively for English and Indian languages, including Marathi, on a total of 1 trillion tokens each. A brief evaluation[14] of these models on Konkani, Gujarati and Marathi MT reveals that they have reasonable translation capabilities via few-shot prompting. We perform greedy decoding without sampling up to 64 new tokens and use chrF for evaluation.

4.2 Results

Table 3 gives the chrF scores[15] for Kadodi-Marathi, Kadodi-English, Marathi-Kadodi and English-Kadodi translation with varying number of shots.

1. Generating Kadodi is challenging: As can be seen, translation into English and Marathi yields better chrF scores than into Kadodi. We found that since the models were not trained on Kadodi, translating into Kadodi leads to very poor translations. In fact, a manual evaluation showed that most of the time the generated translations were in Marathi with some Kadodi word forms. Pronouns and standalone verbs like (is, am, are) are often well handled. In a number of cases for the Sarvam-2b-0.5 model, the Kadodi translations have nothing to do with the sentence being translated, when the source language is English. This is a form of off-target hallucinations. However, Gemma-2-9b and LLaMa-3.1-8b are vastly better. Also note that these models have an easier time handling translation between Marathi and Kadodi compared to translation between English and Kadodi. This is likely because the models have less overhead translating between dialects.

2. Limited impact of shots: Although LLMs are

[13] `https://huggingface.co/sarvamai/sarvam-2b-v0.5`

[14] Since we do not possess any resources for Indo-Portuguese evaluation we skip this but given that Indo-Portuguese is a variation of Portuguese, we expect LLaMa and Gemma to do far better than Sarvam.

[15] nrefs:1 | case:mixed | eff:yes | nc:6 | nw:0 | space:no | version:2.4.1

touted to work well in few-shot settings, even for languages not seen before, we expected that increasing the number of shots would condition the model to better handle Kadodi. For the Sarvam-2b-0.5 model, this is highly translation direction dependent, where Kadodi-Marathi and English-Kadodi generation benefits from increasing shots, but the other two directions barely benefit from shots. On the other hand, LLaMa-3.1-8b and Gemma-2-9b do a significantly better job. Increasing shots from 1 to 4 leads to a large performance jump, but beyond this the gains are minor for up to 12 shots. Comparing Sarvam, Gemma and LLaMa models, it appears that scale indeed is important. Although the latter two models are not intentionally designed for Marathi, they do better and the key difference is the size of the models. Furthermore, the Sarvam model is trained on a vast amount of synthetic data, which might be detrimental.

Since none of the models does particularly well for generating Kadodi, despite our evaluation sentences being simple, we suspect that the reason for this is that they have not seen a shred of Kadodi and even though, it is a dialect of Marathi. They likely consider Kadodi as a garbled version of Marathi. Following the principle of GIGO[16], since the inputs and expected outputs are what the models perceive as noise, the generated content is fairly noisy. This indicates the need for incorporating monolingual Kadodi knowledge into these models, something we leave for future work.

5 Conclusion

In this paper, we presented the first of its kind study of machine translation of Kadodi, a dialect of Marathi spoken in the Vasai region of Maharashtra, India. We described the features of Kadodi and, ***Suman***, a Kadodi-Marathi-English dataset, which was manually created, spanning close to 1,900 tri-parallel entries. Our automatic evaluation showed that Kadodi translation via few-shot prompting of LLMs, even on an Indic exclusive pre-trained language model which as been trained for 1 trillion Indic tokens including Marathi, is still rather poor. This shows that existing LMs, do not handle Kadodi, and likely other dialects of Marathi, indicating a dire situation. However, this means that the field of NLP of Marathi dialects is ripe for exploration. In the future, we would like to expand our dataset, not only to include additional parallel sentences but also branch out to other tasks like summarization, headline generation and question answering, to name a few.

[16]https://en.wikipedia.org/wiki/Garbage_in,_garbage_out

Limitations

This paper focuses on a rather simple case of Kadodi translation, where the resources are small dictionaries and short sentences. However, we plan to scale up data collection and cover more complex sentences spanning multiple domains, subject to annotator availability and budget. We also do not focus on fine-tuning LLMs due to the non-availability of training corpora, but we expect this to be sorted out as our data collection efforts ramp up.

Acknowledgements

Raj Dabre would first like to thank his mother, Mary, who is also the second author, for being an outstanding mother. She carried him as a kid and now hard carries this whole Kadodi NLP project as an annotator and language expert. At the same time, he thanks his honorable father, Noel, and his loving family in general. He also thanks Teresa, the third author, for her enthusiasm and support, along with the entire Kadodi community, which is extremely proud of its language and heritage. Finally, an ode to a great Kadodi person: कहिक्का (kahikka).

References

Kaori Abe, Yuichiroh Matsubayashi, Naoaki Okazaki, and Kentaro Inui. 2018. Multi-dialect neural machine translation and dialectometry. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2022. Hierarchical aggregation of dialectal data for arabic dialect identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4586–4596.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Abhisek Chakrabarty, Raj Dabre, Chenchen Ding, Hideki Tanaka, Masao Utiyama, and Eiichiro Sumita. 2022. FeatureBART: Feature based sequence-to-sequence pre-training for low-resource NMT. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5014–5020, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Abhisek Chakrabarty, Raj Dabre, Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2020. Improving low-resource NMT through relevance based linguistic features incorporation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4263–4274, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Antonio Contarino. 2021. Neural machine translation adaptation and automatic terminology evaluation: a case study on italian and south tyrolean german legal texts.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang

Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

AbdelRahim Elmadany, Sherif Abdou, and Mervat Gheith. 2018. Improving dialogue act classification for spontaneous arabic speech and instant messages at utterance level. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Alexander Erdmann, Nizar Habash, Dima Taji, and Houda Bouamor. 2017. Low resourced machine translation via morpho-syntactic modeling: The case of dialectal Arabic. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 185–200, Nagoya Japan.

Bavtis Dabre Francis Correia, Paul Rumao. 1992. *Samvedi Spoken Language Literature*. Book on Demand.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating similar languages: Evaluations and explorations. In *Proceedings of LREC*.

Jaavid J, Raj Dabre, Aswanth M, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. RomanSetu: Efficiently unlocking multilingual capabilities of large language models via Romanization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593–15615, Bangkok, Thailand. Association for Computational Linguistics.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *Preprint*, arXiv:2401.05632.

Alaa Joukhadar, Huda Saghergy, Leen Kweider, and Nada Ghneim. 2019. Arabic dialogue act recognition for textual chatbot systems. In *Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers*, pages 43–49.

Vijay Keswani and L Elisa Celis. 2021. Dialect diversity in text summarization on twitter. In *Proceedings of the Web Conference 2021*, pages 3802–3814.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, Hans Erik Heje, Ernests Lavrinovics, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, Anders Søgaard, and Johannes Bjerva. 2024. Creoleval: Multilingual multitask benchmarks for creoles. *Preprint*, arXiv:2310.19567.

Zhengyuan Liu, Shikang Ni, Ai Ti Aw, and Nancy F. Chen. 2022. Singlish message paraphrasing: A joint task of creole translation and text normalization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3924–3936, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task. In *Proceedings of the third workshop on NLP for similar languages, varieties and dialects (VarDial3)*, pages 1–14.

Maria das Graças Bruno Marietto, Rafael Varago de Aguiar, Gislene de Oliveira Barbosa, Wagner Tanaka Botelho, Edson Pimentel, Robson dos Santos França, and Vera Lúcia da Silva. 2013. Artificial intelligence markup language: a brief tutorial. *arXiv preprint arXiv:1307.3091*.

Kaushal Kumar Maurya, Rahul Kejriwal, Maunendra Sankar Desarkar, and Anoop Kunchukuttan. 2023. Utilizing lexical similarity to enable zero-shot machine translation for extremely low-resource languages. *arXiv preprint arXiv:2305.05214*.

Olubusayo Olabisi, Aaron Hudson, Antonie Jetter, and Ameeta Agrawal. 2022. Analyzing the dialect diversity in multi-document summaries. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6208–6221.

Teresa Pereira. 2007. *Samvedi Community*. Book on Demand.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Stutzman, Bismarck Odoom, Sanjeev Khudanpur, Stephen Richardson, and Kenton Murray. 2024. Kreyòl-MT: Building MT for Latin American, Caribbean and colonial African creole languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3083–3110, Mexico City, Mexico. Association for Computational Linguistics.

J. Russell and R. Cohn. 2012. *Kadodi Language*. Book on Demand.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh

Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the dsl shared task 2014. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, pages 58–67.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.