A Survey of NLP Progress in Sino-Tibetan Low-Resource Languages

Shuheng Liu and **Michael Best**
Georgia Institute of Technology
sliu775@gatech.edu, mikeb@gatech.edu

Abstract

Despite the increasing effort in including more low-resource languages in NLP/CL development, most of the world's languages are still absent. In this paper, we take the example of the Sino-Tibetan language family which consists of hundreds of low-resource languages, and we look at the representation of these low-resource languages in papers archived on ACL Anthology. Our findings indicate that while more techniques and discussions on more languages are present in more publication venues over the years, the overall focus on this language family has been minimal. The lack of attention might be owing to the small number of native speakers and governmental support of these languages. The current development of large language models, albeit successful in a few quintessential rich-resource languages, are still trailing when tackling these low-resource languages. Our paper calls for the attention in NLP/CL research on the inclusion of low-resource languages, especially as increasing resources are poured into the development of data-driven language models.

1 Introduction

Research on low-resource languages (LRLs) has become a major topic in the Natural Language Processing (NLP) and Computational Linguistics (CL) as seen by promotions from various conferences and workshops (Magueresse et al., 2020), especially as we strive to develop multilingual large language models (LLMs) to incorporate more languages (Mani and Namomsa, 2023). ChatGPT (OpenAI et al., 2024), one of the most successful commercially available LLMs, is claimed to support more than 80 languages (Funelas, 2024). However, this number is still far lower than the total number of languages in the world, which is estimated to be more than 7,000 (Eberhard et al., 2024). As Joshi et al. (2020) pointed out, we would want to improve the linguistic diversity and inclusion in NLP/CL development, as "unsupervised pre-train methods only make the 'poor poorer'."

One common way to improve the performance of language models on LRLs is cross-lingual transfer learning (Conneau et al., 2018), with which tasks in a target language are evaluated by models trained on source language(s). This technique has been found to work better when the source languages include languages from the same family as the target language (Pandya and Bhatt, 2024; Woller et al., 2021). A language family is a group of "genetically" related languages which all descended from the same ancestral language (Nádasdy, 1993). In this paper, we focus on the Sino-Tibetan (ST) language family, with the fourth largest number of languages (Eberhard et al., 2024) and the second largest number of speakers (Handel, 2008). It consists of Sinitic languages such as Mandarin and Cantonese, and Tibeto-Burman languages Tibetan and Burmese, that descended from a hypothetical common ancestor Proto-Sino-Tibetan. There are over 1 billion native speakers of ST languages, mostly occupying "East Asia, peninsular Southeast Asia, and parts of South Asia" (Handel, 2008). A figure of the geographic distribution of Sino-Tibetan languages is shown in Appendix A.

Chinese[1] has been the most focused ST languages by the NLP family such that the late ChatGPT is able to perform well on natural language tasks (Fang et al., 2023; Li et al., 2023) and ERNIE bot (Baidu Research), a Chinese LLM, was developed not long after the release of ChatGPT. However, it is yet unknown how other ST languages are represented in the NLP literature, especially since they are all considered as LRLs.

Our study attempts to answer on the following research questions:

[1]The term "Chinese" in the context of NLP usually refers to Mandarin in its written form.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7804–7825

RQ1. What is the overall and individual coverage of low-resource ST languages in NLP/CL literature?

RQ2. What are the main topics of these NLP/CL publications, and how have they changed over the years?

RQ3. What is the publication trend for ST languages in various NLP/CL conferences?

We focus on the papers archived in ACL Anthology, which contains publications from well-known NLP/CL conferences. We limit our study to include the low-resource ST languages that are spoken by more than 100,000 people, which are more likely to appear in NLP/CL literature. Our findings show that only 0.35% of all publications cover these ST languages. Less than half of the languages selected are covered, and more than 80% of the papers focus on just 5 languages. While it seems that the objective topics (aspects and methods of NLP/CL) covered by these publications have increased over the years, most publications have been and are still centered around 4 topics. Lastly, while we are excited to see the increasing number of publications in various conferences over the years, 81.34% of the papers are published in venues that draw less attention from the research community.

We start our paper with a brief literature review in § 2, followed by our methodology in § 3. We present our results in § 4, and discuss their relevance in § 5. Our paper acknowledges the progress of NLP/CL in ST LRLs, and simultaneously highlights areas for further development. We hope that our study fills the gap in this line of literature, calling attention to overlooked LRLs, and encouraging NLP/CL development with increasing diversity.

2 Literature Review

2.1 NLP for LRLs

Joshi et al. (2020) categorized a language into one of the 6 classes, from class 0 (The Left-Behinds) to class 5 (The Winners), based on the availability of labeled and unlabeled data of that language. All languages in class 0, 1, 2, 3, and occasionally some languages in 4, appear as LRLs in various literature. While newer language models, especially multilingual models (Devlin et al., 2019; Conneau et al., 2020) and large language models (Workshop et al., 2023; OpenAI et al., 2024; Touvron et al., 2023), attempt to include languages from class 4, 3 and even 2 sometimes, the number of languages in class 0 and 1 still take up of more than 90% of world's languages.

To improve the model performance on LRLs tasks, there are usually two approaches. A direct approach would be to increase the amount of resource for LRLs, i.e. create more data for these languages. When there are unlabeled data for a language, researchers would be to create labeled data, either manually (Nivre et al., 2020; Mayhew et al., 2024) or heuristically (Pan et al., 2017; Wang et al., 2021). However, creating labeled data, especially manually, is known to be cost-ineffective (Kang et al., 2023). Hedderich et al. (2021) provided an overview on the potential algorithmic and engineering solutions for developing NLP technology in low-resource settings. These include methods to create artificial data such as data augmentation (Evuru et al., 2024; Lucas et al., 2024; Sobrevilla Cabezudo et al., 2024), and methods to effectively utilize the available resources such as cross-lingual transfer learning (Conneau et al., 2018; Ruder et al., 2019), showing remarkable improvement on low-resource languages. In recent years, there has been more reliance on the power of LLMs due to their capability in zero-shot cross-lingual transfer (Artetxe and Schwenk, 2019) when all they need is unlabeled data.

2.2 Survey Papers on Language Groups

Despite having the classification of language family, surveys often group languages based on their geographical distribution. For example, Zhang et al. (2024) presents a survey on neural machine translation for LRLs in China, albeit focusing on non-ST languages such as Mongolian and Uyghur. Other country-wide language studies include NLP progress for Ghanaian languages (Azunre et al., 2021; Issaka et al., 2024) and Indian languages (Khanuja et al., 2023; Vijayvergia et al., 2023), countries that are linguistically diverse with many LRLs. Such research can sometimes escalate to a continental level, such as for African languages (Adebara and Abdul-Mageed, 2022) and Latin American languages (Tonja et al., 2024). While Gopal and Haroon (2016) attempts to discuss the Dravidian languages in general, the scope was limited only to 4 languages with the most speakers, although this was possibly due to the availability of language technologies being limited to just these 4 languages.

Language families are determined based on regularity hypothesis, that is, languages are likely to be derived from the same parent language if there are numerous similarities (Rowe and Levine, 2015). Such similarities can still be detectable (Kumar et al., 2021), which can potentially be exploited for various language modeling techniques. It is therefore imperative to understand what the research status of language technology on various language families is and address gaps in current research. To the best of our knowledge, there has not been any survey on the development of NLP/CL on the Sino-Tibetan language family as a whole. We intend to offer a first glance at this problem, and hope to draw more attention to this particular language family, LRLs, and language family studies.

3 Method

3.1 Data Collection and Querying

Our study focuses on low-resource ST languages with more than 100,000 speakers reported by the World Atlas of Language Structures (WALS) Online (Dryer and Haspelmath, 2013) under the Creative Commons Attribution 4.0 International License (CC BY 4.0), yielding 49 languages in total. As researchers sometimes use different names of the same language, e.g. "Bodo" and "Boro" for the Bodo language, or "Hokkien", "Southern Min" and "Min Nan" for the Southern Min language, we constructed a list of alternative language names to find relevant papers. Each language therefore contains a list of n query terms $q_1, q_2, \cdots, q_n$. Some language names such as Bai coincide with common Asian surnames that could appear as references in a paper, and we added the word "language" to prevent querying thousands of irrelevant papers. The list of all languages and the names used can be found in Appendix B.

We focused on the papers available in ACL Anthology, which archives papers published in well-known international NLP and CL conferences as well as NLP conferences that have a regional focus. The ACL OCL corpus (Rohatgi et al., 2023) contains extracted full text of 73,285 papers in ACL Anthology from 1952 to 2022 under the Creatie Commons Attribution-Noncommercial 4.0 International License (CC BY-NC 4.0).

For each language, we used regex string match to match papers in which one of the query terms of the language appeared in the ACL OCL corpus. The regex query that we used is

$$\texttt{\textbackslash b}q_1\texttt{\textbackslash b}|\texttt{\textbackslash b}q_2\texttt{\textbackslash b}|\cdots|\texttt{\textbackslash b}q_n\texttt{\textbackslash b}$$

where each query term is surrounded by the regex metacharacter \b and joined with the | character. We also printed out the 200 characters before and after each matched query term to provide context for the next step. We repeated this for all 49 languages, resulting in 1,124 matched papers.

3.2 Data Cleaning and Annotation

The previous step ensured that each matched paper contained a query term. However, the mention did not imply that the paper was building NLP tools for or conducting linguistic analysis of that language. For example, Tiedemann and Nakov (2013) wrote "The first step in our pivoting experiments involves SMT between closely related languages..., e.g. ..., **Cantonese**-Mandarin" which mentioned Cantonese as a background reference rather than built a translation tool for the language. Some query terms might coincide with mentions that are not about the corresponding languages. For example, Okabe et al. (2022) wrote "Japhug, a language from the Sino-**Tibetan** family" and was matched for the Tibetan language.

Using the 400-character context from the previous step, the first author examined all the matched papers only to retain papers relevant to our study of NLP/CL progress in ST languages (ST papers). Whenever the context is not enough to decide if a paper is relevant, the author would look at the original PDF of the paper. Eventually, we only retained 274 relevant matches (24.38% of regex-matched papers). We randomly sampled 100 papers from the rest of the ACL OCL corpus and found that none of them were relevant to our study, suggesting that there are unlikely any false negatives.

Rohatgi et al. (2023) defined 21 *objective topics* based on the submission topics in major CL conferences, with which the author labeled papers. These objective topics are used as labels for the first author to assign to each paper. Each paper received at least one label based on its content. Eventually 16 of the 21 topics appeared in our dataset, namely dialogue and interactive systems (**Dialogue**); generation (**NLG**); information extraction (**IE**); interpretability and analysis of models for NLP (**Interpret**); linguistic theories, cognitive modeling and psycholinguistics (**LingTheory**); machine learning for NLP (**ML**); machine translation and multilinguality (**MT**); phonology, morphology and word segmentation (**WS**); question answering (**QA**); resources and evaluation (**Resource**); semantics: lexical semantics (**LexSem**); semantics: sentence-level semantics, textual inference (**SenSem**); sentiment analysis, stylistic analysis, and argument mining (**Sentiment**); speech and multimodality (**Speech**); summarization (**Summ**) and syntax: tagging, chunking and parsing (**Syntax**).

4 Results

4.1 ST Languages in ACL Anthology

Out of the 73K papers in the ACL OCL corpus, we only found 274 matches, i.e. 274 instances where a paper is discussing NLP or CL progress of one of the target languages. As each paper can be about more than one language, there are 253 unique ST papers (0.35%) found in our study.

Table 1 shows the distribution of the number of papers found for each language. More than half (26 out of 49, 53.06%) of our target languages did not yield any matches and 18 (36.73%) yielded fewer than 10 matches. Only five languages (10.20%), namely Burmese (57 matches), Cantonese (96 matches), Hakka (16 matches), Southern Min (17 matches) and Modern Literary Tibetan (44 matches), exceeded 10 matches, and they take up 83.94% of all the matches. This list highly correlates with the language taxonomy defined in Joshi et al. (2020). These five languages, plus Fuzhou and Wu, are the only 7 class 1 languages ("with some unlabeled data" and "have almost no labeled data"), whereas all other languages are class 0 languages ("with exceptionally limited resources" and "ignored from the perspective of language technologies"). The presence of 17 class 0 languages in Table 1 is encouraging, but it corroborates with the claim that these languages, along with those that are absent from this table, are largely overlooked by the NLP/CL community.

Moreover, Schwartz (2022) looked at 9,602 ACL abstracts between 2013 and 2021 and found that only 432 (4.50%) mentioned at least one language from the ST language family. This statistics would most likely be even lower when we leave out Chinese. This observation, combined with our finding of fewer than 300 ST papers, shows that despite being the 4th largest language family by the number of languages, low-resource ST languages are disproportionately represented in ACL Anthology.

4.2 NLP Methods for ST Languages

Figure 1 shows the distribution of papers in each topic in each year. **LingTheory**, **Resource**, **MT** and **Speech** have almost always been focused on

# Papers	Languages
< 10	Ao, Arakanese (Marma), Bodo, Dimasa, Garo, Jingpho, Karen (Sgaw), Kok Borok, Lahu, Limbu, Lotha, Meithei, Mikir, Mizo, Naxi, Newari (Kathmandu), Tamang (Eastern), Wu
$10 - 49$	Hakka, Min (Southern), Tibetan (Modern Literary)
$50 - 100$	Burmese, Cantonese

Table 1: Number of papers found for each language. The left column is the range of the number of papers found, and the right column is the lists of languages.

through the years, and more papers are being published on these topics as time progresses. We can also see a trend of more papers being published over the years, and more topics are covered by these papers in more recent years. However, despite spanning over 16 topics, more than half of the topics were only minimally covered, shown by the sporadic light cells in the right half of Figure 1.

While it is comforting to see that the CL objective topics covered for ST languages have increased over the years, Figure 2 re-emphasizes how the papers are disproportionately distributed towards the top languages. Although there is almost always at least one **LingTheory** paper for each language, the non-zero cells become more sparse as we move further to the right half of the figure. The diversity of the objective topics is, once again, mostly contributed by the top five languages (Burmese, Cantonese, Hakka, Southern Min and Modern Literary Tibetan), of which papers with the objective topics **SenSem**, **IE**, **Sentiment**, **Summ**, **Dialogue**, **NLG** and **QA** can only be found.

What we found in these two subsections is encouraging and alarming at the same time. On the one hand, Figure 1 shows a good prospect of more papers on low-resource ST languages covering a wider range of topics being published in the future. On the other hand, most papers cover, and most topics are discussed only for a few relatively higher-resource languages.

4.3 Paper Distribution in Conferences

ACL Anthology collects papers from both ACL events as well as non-ACL events that publish CL papers. We then analyzed distribution of our dataset in these events and the changes from 2000

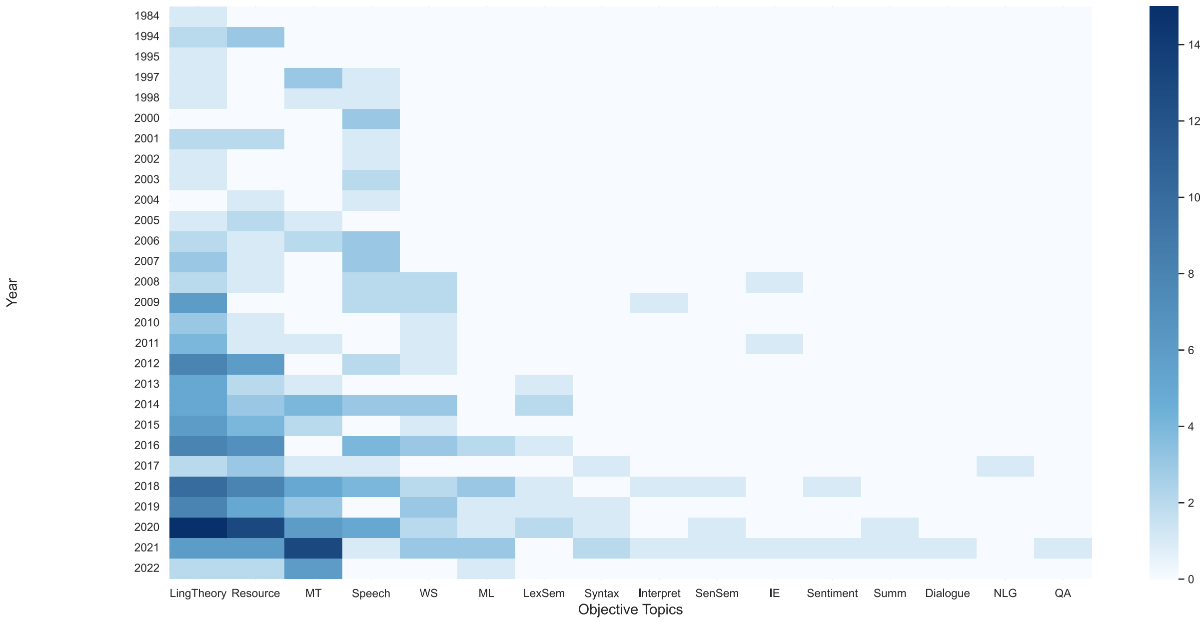

Figure 1: Heat map showing the distribution of papers in each objective topic in each year. Darker cells represents more papers in a topic and year. Objective topics are sorted by the total number of papers, with the highest on the left and lowest on the right.

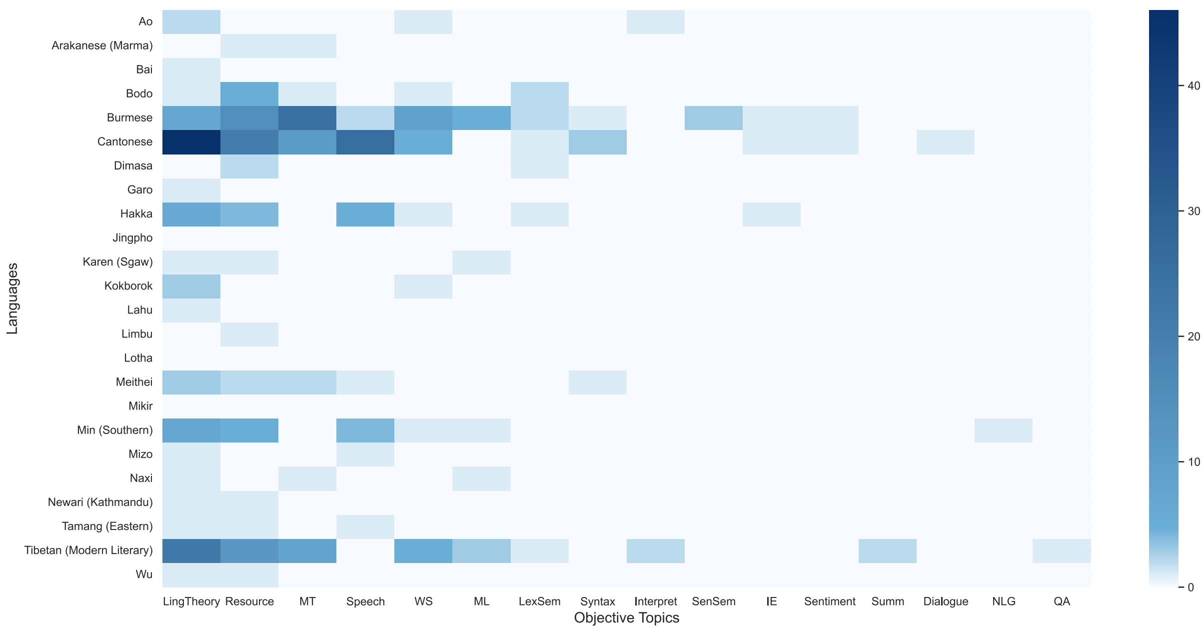

Figure 2: Heat map showing the distribution of papers in each topic for each language. Darker cells represents more papers in a topic and year. Objective topics are ordered by the total number of papers, with the highest on the left and lowest on the right.

to 2021. We excluded workshop papers, and omitted conferences with not more than five papers in our dataset. 134 ST papers across 9 conferences, namely ACL, IJCNLP, LREC, EMNLP, CCL, PACLIC, COLING, ROCLING and IJCLCLP, satisfy the criteria above (full names of these conferences are provided in Appendix C). As IJCNLP, LREC and COLING are held every two years, we put the papers into two-year buckets to capture the general trend. We also calculated the percentages of all ST papers published in each two-year bucket. The results are shown in Figure 3.

We observe from Figure 3 that as the number of publications increases over the years, these publications are also distributed in a more diverse range of conferences, from only 3 in 2000-2001 to 8 in 2020-2021. PACLIC, LREC and ACL comprise of 55.22% of all papers. The number of paper published in EMNLP, being the third most impactful conference hosted on ACL Anthology (Eickhoff, 2023), is observed to be growing, as well as CCL, the largest NLP conference in China focusing on languages in China,[2] whose papers are only featured in ACL Anthology after 2020.

The percentage of ST papers mostly fluctuated between 0.20% to 0.30%, but the percentage has increased to 0.36% in the most recent 2020-2021 bucket. Despite the low percentage, the general upward trend gives us some confidence that the inclusion of ST languages is growing.

Nevertheless, ACL and EMNLP are the only ACL events in this list, while 81.34% of the papers are published in non-ACL events. The h5-index[3] of ACL is 192 and 176 for EMNLP. COLING (73), LREC (61) and IJCNLP (24) have a significantly lower h5-index, and those for the other four conferences were not found.[4] We are not suggesting that the h5-index informs the quality of the publication venues, but it can deduce that these publications have not received enough attention.

5 Discussion

5.1 NLP Research in ST Languages

In contrast to the increasing recognition and focus in the study of multilingual NLP and LRLs, exemplified by the inclusion of the "Linguistic Diversity" and "Multilingualism and Cross-lingual NLP" tracks in the recent ACL 2023 (Rogers et al., 2023), our results in Section 4.1 show that ST languages have been disproportionately represented in the NLP literature. Even when we look at the languages covered in multilingual corpora containing hundreds of thousands of languages, only a few ST languages are usually included. Table 2 presents a few highly cited multilingual corpora and their coverage on ST languages. From Table 2, we can see that in most corpora, only $\leq 5\%$ languages are in the ST family, including Simplified/Traditional Chinese which is always present. Similar to Table 1, Burmese, Cantonese (Yue Chinese) and Tibetan are the next most popular choices. Even when some other LRLs are included, they usually make up of a small percentage of the entire dataset, such as in WikiAnn (Pan et al., 2017) named entity recognition (NER) dataset, where all ST languages other than Chinese (Mandarin) amount to 131K name mentions, slightly more than the name mentions from Vietnamese alone (125K), and without Southern Min (Chinese (Min Nan)) and Cantonese, this number drops to only 31K.

So what makes Burmese, Cantonese, Tibetan, Southern Min and Hakka from Table 1 different from the other ST LRLs? We believe that there are two explanations to this question, in addition to the fact that they are class 1 languages.

The first explanation is the number of native speakers of these languages. Cantonese, Southern Min, Hakka and Burmese have the 1st, 3rd, 4th and 5th most speakers of all low-resource ST languages (Dryer and Haspelmath, 2013). While Central Tibetan has the 15th most speakers, this number might have only included the Ü-Tsang (Central) Tibetan speakers. As the three main dialects of Tibetan (or the three main Tibetic languages, i.e. Ü-Tsang, Khams and Amdo Tibetan) are not used as written languages, and Classical Literary Tibetan has been and still is widely used for writing (Tournadre, 2014), the number of speakers (or users) might be much higher when the language name is only specified as "Tibetan" in these corpora. Simply adding the speaker population of these three dialects raises the ranking of Tibetan to the 7th.

The absence of the 2nd (Wu) and 6th (Fuzhou) languages with the most speakers, which are also the class 1 languages, brings out the second explanation to this question - the lack of support and recognition of these languages as individual "languages". Burmese is the official language of Myanmar and even though Cantonese (Bolton, 2011),

[2] http://cips-cl.org/static/CCL2024/en/index.html

[3] h5-index is the h-index for articles published in the last 5 complete years.

[4] Data from Google Scholar and Research.com

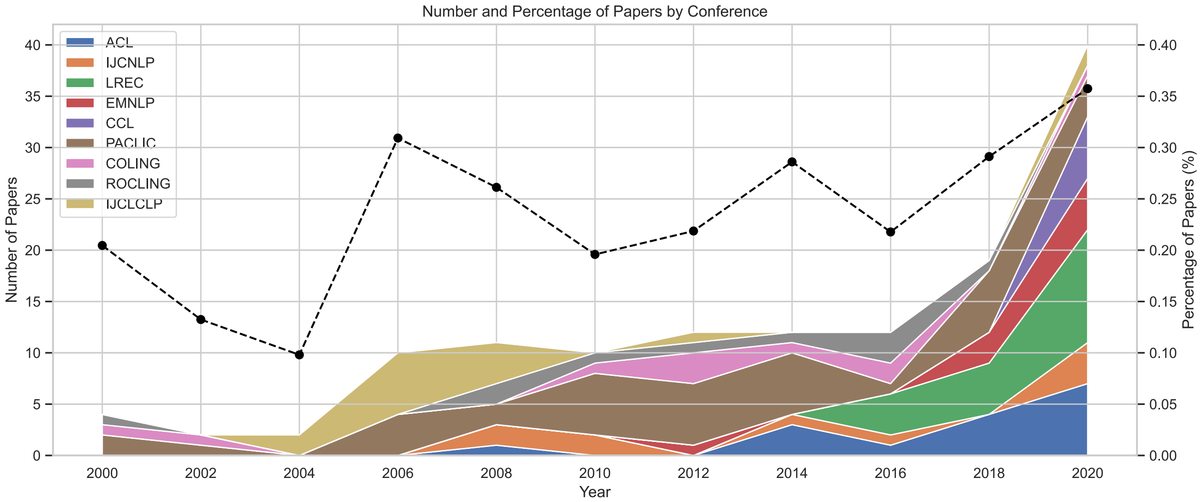

Figure 3: Number of papers about ST languages (ST papers) in different conferences between 2000 and 2021. The papers are grouped into two-year buckets as IJCNLP, LREC and COLING are held biannually. The width of the band of each color represents the number of papers in the corresponding conference in each two-year bucket. The dashed line plot shows the trend of the percentage of ST papers published in two-year buckets.

Corpus	# Langs	ST Languages
FLORES-200 (Team et al., 2022)	204	Standard Tibetan, Dzongkha, Jingpho, Mizo, Meitei, Burmese, Yue Chinese, Chinese
OSCAR (Abadji et al., 2022)	153	Burmese, Chinese, Newari, Tibetan, Wu Chinese
ROOTS (Laurençon et al., 2023)	46	Simplified Chinese, Traditional Chinese
UD (Nivre et al., 2020)	90	Cantonese, Chinese, Classical Chinese
WikiAnn (Pan et al., 2017)	282	Tibetan, Min Dong Chinese, Dzongkha, Gan Chinese, Hakka Chinese, Sichuan Yi, Burmese, Newari, Wu Chinese, Classical Chinese, Chinese (Min Nan), Cantonese, Chinese
XTREME (Hu et al., 2020)	40	Burmese, Mandarin

Table 2: Examples of large multilingual corpora with the total number of languages and the ST languages they cover. The names of the languages appear as how they are defined in the corresponding corpus.

Southern Min and Hakka (Chen, 2020), and Tibetan (Wan and Zhang, 2007) are not official languages of any state, they have received various level of governmental support and attention. In contrast, Wu and Fuzhou are widely recognized as simply "Chinese dialects", which is justified politically and culturally, but not linguistically (Handel, 2015). Such lack of recognition as individual languages could result in NLP/CL researchers without training in Linguistics to ignore the development of technology for these languages. Even with researchers working on these languages, speakers would usually engage in verbal communication in these languages, instead of written (Mair, 2013), causing a lack of textual data available for the NLP community. Additionally, despite some, if not all, non-Sinitic languages being recognized by local authorities, many (in China as an example) are facing the lack of specific legislation for minority language planning (Sun, 2015). Many areas in which Tibeto-Burman languages are spoken are also challenging for researchers to access and collect data (Matisoff, 2015). These might hinder the development of the languages themselves, as well as the NLP/CL research on these languages.

5.2 LRLs in the LLM Era

Since the release of ChatGPT in 2022[5], there have been more and more large language models (LLMs) developed (Touvron et al., 2023; Team et al., 2024). The performances of these LLMs rely heavily on

[5] https://openai.com/index/chatgpt/

the amount of textual data available (Hoffmann et al., 2024) to be proficient in multiple languages. Despite the high performance of LLMs on major languages in the world, that on LRLs is still trailing (Avetisyan and Broneske, 2023; Robinson et al., 2023; Adelani et al., 2024).

We therefore conducted an additional experiment by running ChatGPT (*gpt-3.5-turbo*) on the low-resource ST languages from the WikiAnn NER dataset (Pan et al., 2017).[6] The method of the experiment can be found in Appendix D, and we record the micro F_1 scores in Table 3.

From Table 3, we observe that the F_1 scores for the Tibeto-Burman languages, i.e. Burmese (4.13%), Dzongkha (12.12%), Newari (9.39%), Tibetan (3.60%) are extremely low. The F_1 scores of 4 of the 6 Sinitic languages, Cantonese (7.06%), Gan Chinese (17.82%), Hakka (29.18%), Min Dong Chinese (20.41%), are lower than the performance on Weibo (30.09%), the lowest from Xie et al. (2023). This resonates with the results from previous findings that LLMs are trailing behind on LRLs. Additionally, even the highest F_1 score in this table is merely 35.97%. While ChatGPT has been shown not to perform on par with the state-of-the-art fine-tuning models on the commonly used English NER dataset (Tjong Kim Sang and De Meulder, 2003) at only 53.5% (Qin et al., 2023), the numbers shown in Table 3 are still much lower. This highlights the gap between ChatGPT's performance on high- and low-resource languages.

The *curse of multilinguality*, i.e. degradation of multilingual models on individual languages due to model capacity (Conneau et al., 2020), is known to harm the performance of language models on LRLs (Wu and Dredze, 2020). While methods to mitigate such phenomenon have been proposed (Pfeiffer et al., 2022; Blevins et al., 2024), we are yet to know if LLMs such as ChatGPT have incorporated these methods, or if these methods can be generalized to trillion-parameter models.

5.3 Future Directions and Recommendations

One of the first challenges that we encountered was the difficulty to find natural language processing papers on specific natural languages. Ducel et al. (2022) found that publications in LREC tend to respect the #BenderRule ("Always name the language(s) you're working on") (Bender, 2021)

[6]There are no data for Sichuan Yi despite the language appearing in the paper.

[7]`https://catalog.ldc.upenn.edu/LDC2011T03`

Language	F_1 (%)
Tibetan	3.60
Burmese	4.13
Cantonese*	7.06
Newari*	9.39
Dzongkha	12.12
Gan Chinese	17.82
Min Dong Chinese	20.41
Hakka	29.18
Chinese (Min Nan)*	33.94
Wu Chinese	35.97
Weibo (Peng and Dredze, 2015)	30.09
Onto. 4[7]	33.74
MSRA (Zhang et al., 2006)	45.51

Table 3: F_1 scores of ChatGPT on the low-resource ST languages in WikiAnn in ascending order. The three results from the bottom are the performance reported by Xie et al. (2023) on three general-domain Chinese datasets using the vanilla zero-shot method. *Only the first 3,000 sentences were used for evaluation.

more than those in ACL. Even when languages are cited in the publications, there can be more efficient ways developed to query these publications, such as adding keywords to make the searches easier, or have the authors indicate clearly the language(s) at which their papers are targeting as part of the metadata of the paper.

Our paper, while only focusing on the ST language family, hopes to call for more NLP/CL research on the effect of using languages from the same family. In addition to cross-lingual transfer (Pandya and Bhatt, 2024; Woller et al., 2021), language family has also shown to be helpful for data augmentation (Scalvini and Debess, 2024). Moreover, as languages in the same family often share cognates and regular sound changes, such features can potentially be utilized by language models to generalize to LRLs in the same family as higher-resource languages, with multimodal language models or language models trained with phonetic/phonological knowledge. The NLP/CL progress on language family might even reciprocally benefit the historical linguists in their determination of language families (Kondrak, 2009).

While most NLP research on LRLs works under the low-resource restriction, many tried to tackle the problem by creating more resource for these languages. More recently, effort has been put into groups of LRLs, such as African languages by

Masakha NLP (Adelani et al., 2021; Dione et al., 2023; Adelani et al., 2023) and Southeast Asian languages (Lovenia et al., 2024). From Figure 2, despite **Resource** consisting the second most papers, not only does this topic also contains evaluation papers, but they are also highly concentrated on a few languages. An example of such lack of unlabeled data is the distribution of languages[8] for CommonCrawl[9], which one of the widely used corpora collected from the Internet for training language models. For ST languages, the CommonCrawl corpus only contains the Chinese languages (not specified which ones), Tibetan, Burmese and Dzongkha. This lack of unlabeled textual data alludes to the lack of online discourses in these LRLs, and potentially the lack of online communities of speakers of these LRLs. A future direction for researchers can be creating more resources for these languages, even just unlabeled data. In addition to collecting data, researchers may think of ways to foster online communities for LRL speakers.

6 Conclusion

Our paper is the first in the field to provide a systematic look into the coverage of ST languages in NLP/CL literature. From our analyses, we acknowledge the encouraging improvement made by the NLP/CL community to include more ST languages over the years, and have a wide range of techniques and discussion topics about these languages. However, we also raise concerns on the low count of publications, covering only part of our list of languages, despite the list having already excluded hundreds of ST languages with even lower number of native speakers. The publication venues have also become more diverse, but most papers possibly did not receive enough attention from the research community. The existing disparity of the literature coverage and model performance on ST LRLs as compared to high-resource languages may be worsened by the advancement of LLMs, relying on large quantity of unlabeled data, making the "poor poorer" (Joshi et al., 2020). We hope that this paper calls for more research into ST languages and ways to improve the status of LRLs.

[8] `https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html`, accessed on Nov. 23rd, 2024.

[9] `https://commoncrawl.org`

Limitations

The queries we used in Section 3.1 to reduce false positives might still overshoot to exclude publications that should be in our dataset, even after manually checking a random sample. One reviewer pointed out that a lot of metadata of submissions, including the languages studied, are collected but not released for publications. These metadata could allow future analysis similar to this paper to provide insights on how the field of NLP/CL can advance. We hope that in the future, conferences can include the language(s) that are studied in each paper. Additionally, we would like to call for NLP/CL researchers to clarify the language(s) that they studied in their publications.

One reviewer pointed out that there could be some more fine-grained classification provided in Section 4.2 to generate more insights on the research bias. We agree with the reviewer and attempted a small-scale analysis, but still decide to leave more rigorous analyses for future work.

While ACL Anthology collects publications in major NLP events, NLP literature, especially regarding indigenous ST languages, can be published in other venues. Additionally, almost all of the papers in our dataset are published in English, while there can be publications in other languages. Future work may explore literature in other Computer Science/Computational Linguistics venues as well as in other languages, especially Chinese, Burmese and various official languages in India.

Acknowledgments

We thank the three anonymous reviewers for their recognition and valuable feedback. Our gratitude also extends to Karthik S. Bhat, Terra Blevins, Amy Z. Chen, Zev Handel, Aman Khullar and Daniel Nkemelu for their support and discussions on various ideas and aspects of this paper. We would also like to thank the rest of the Georgia Tech T+ID lab for their feedback on earlier versions of this paper.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. *Preprint*, arXiv:2201.06642.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, A. Seza Doğruöz, André Coneglian, and Atul Kr. Ojha. 2024. Comparing LLM prompting with cross-lingual transfer performance on indigenous and low-resource Brazilian languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 34–41, Mexico City, Mexico. Association for Computational Linguistics.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Hayastan Avetisyan and David Broneske. 2023. Large language models and low-resource languages: An examination of Armenian NLP. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 199–210, Nusa Dua, Bali. Association for Computational Linguistics.

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah. 2021. Nlp for ghanaian languages. *Preprint*, arXiv:2103.15475.

Baidu Research. 2023. Introducing ernie 3.5: Baidu's knowledge-enhanced foundation model takes a giant leap forward.

Emily M. Bender. 2021. The #benderrule: On naming the languages we study and why it matters.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10822–10837, Miami, Florida, USA. Association for Computational Linguistics.

Kingsley Bolton. 2011. Language policy and planning in hong kong: Colonial and post-colonial perspectives. *Applied Linguistics Review*, 2(2011):51–74.

Suchiao Chen. 2020. Language policy and practice in taiwan in the early twenty-first century. In *Language Diversity in the Sinophone World*, pages 122–141. Routledge.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

Fanny Ducel, Karën Fort, Gaël Lejeune, and Yves Lepage. 2022. Do we name the languages we study? the #BenderRule in LREC and ACL articles. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 564–573, Marseille, France. European Language Resources Association.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World. Twenty-seventh edition.* SIL International. Online version: `http://www.ethnologue.com`.

Carsten Eickhoff. 2023. Impact factors for computer science conferences. *Preprint*, arXiv:2310.08037.

Chandra Kiran Evuru, Sreyan Ghosh, Sonal Kumar, Ramaneswaran S, Utkarsh Tyagi, and Dinesh Manocha. 2024. CoDa: Constrained generation based data augmentation for low-resource NLP. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3754–3769, Mexico City, Mexico. Association for Computational Linguistics.

Changchang Fang, Yuting Wu, Wanying Fu, Jitao Ling, Yue Wang, Xiaolin Liu, Yuan Jiang, Yifan Wu, Yixuan Chen, Jing Zhou, Zhichen Zhu, Zhiwei Yan, Peng Yu, and Xiao Liu. 2023. How does ChatGPT-4 preform on non-english national medical licensing examination? an evaluation in chinese language. *PLOS Digit. Health*, 2(12):e0000397.

Raphaella Funelas. 2024. Chatgpt language capabilities.

Sreelakshmi Gopal and Rosna P. Haroon. 2016. Word sense disambiguation on dravidian languages: A survey. In *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, ICTCS '16, New York, NY, USA. Association for Computing Machinery.

Zev Handel. 2008. What is sino-tibetan? snapshot of a field and a language family in flux. *Language and Linguistics Compass*, 2(3):422–441.

Zev Handel. 2015. 34the classification of chinese: Sinitic (the chinese language family). In *The Oxford Handbook of Chinese Linguistics*. Oxford University Press.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2024. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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. *Preprint*, arXiv:2003.11080.

Sheriff Issaka, Zhaoyi Zhang, Mihir Heda, Keyi Wang, Yinka Ajibola, Ryan DeMar, and Xuefeng Du. 2024. The ghanaian nlp landscape: A first look. *Preprint*, arXiv:2405.06818.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Junmo Kang, Wei Xu, and Alan Ritter. 2023. Distill or annotate? cost-efficient fine-tuning of compact models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11100–11119, Toronto, Canada. Association for Computational Linguistics.

Simran Khanuja, Sebastian Ruder, and Partha Talukdar. 2023. Evaluating the diversity, equity, and inclusion of NLP technology: A case study for Indian languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1763–1777, Dubrovnik, Croatia. Association for Computational Linguistics.

Grzegorz Kondrak. 2009. Identification of cognates and recurrent sound correspondences in word lists. In *Traitement Automatique des Langues, Volume 50, Numéro 2 : Langues anciennes [Ancient Languages]*, pages 201–235, France. ATALA (Association pour le Traitement Automatique des Langues).

Sourav Kumar, Salil Aggarwal, Dipti Misra Sharma, and Radhika Mamidi. 2021. How do different factors impact the inter-language similarity? a case study on Indian languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 112–118, Online. Association for Computational Linguistics.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2023. The bigscience roots corpus: A 1.6tb composite multilingual dataset. *Preprint*, arXiv:2303.03915.

Linhan Li, Huaping Zhang, Chunjin Li, Haowen You, and Wenyao Cui. 2023. Evaluation on ChatGPT for Chinese Language Understanding. *Data Intelligence*, 5(4):885–903.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James Validad Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Jann Railey Montalan, Ryan Ignatius Hadiwijaya, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus Irawan, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johanes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Tai Ngee Chia, Ayu Purwarianti, Sebastian Ruder, William Chandra Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng Xin Yong, and Samuel Cahyawijaya. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Agustín Lucas, Alexis Baladón, Victoria Pardiñas, Marvin Agüero-Torales, Santiago Góngora, and Luis Chiruzzo. 2024. Grammar-based data augmentation for low-resource languages: The case of Guarani-Spanish neural machine translation. 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 6385–6397, Mexico City, Mexico. Association for Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *Preprint*, arXiv:2006.07264.

Victor H Mair. 2013. The classification of sinitic languages: What is 'chinese'. *Breaking down the barriers: Interdisciplinary studies in Chinese linguistics and beyond*, pages 735–754.

Ganesh Mani and Galane Basha Namomsa. 2023. Large language models (llms): Representation matters, low-resource languages and multi-modal architecture. In *2023 IEEE AFRICON*, pages 1–6.

James A. Matisoff. 2015. The sino-tibetan language family.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Suppa, Hila Gonen, Joseph Marvin Imperial, Börje Karlsson, Peiqin Lin, Nikola Ljubešić, Lester James Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. 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 4322–4337, Mexico City, Mexico. Association for Computational Linguistics.

Á. Nádasdy. 1993. Language families. In *Language and Speech*, pages 14–16, Vienna. Springer Vienna.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Shu Okabe, Laurent Besacier, and François Yvon. 2022. Weakly supervised word segmentation for computational language documentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7385–7398, Dublin, Ireland. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Hariom A. Pandya and Brijesh S. Bhatt. 2024. Does learning from language family help? a case study on a low-resource question-answering task. *Natural Language Processing*, page 1–18.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for Chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554, Lisbon, Portugal. Association for Computational Linguistics.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Anna Rogers, Marzena Karpinska, Jordan Boyd-Graber, and Naoaki Okazaki. 2023. Program chairs' report on peer review at acl 2023. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages xl–lxxv, Toronto, Canada. Association for Computational Linguistics.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: Advancing open science in computational linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361, Singapore. Association for Computational Linguistics.

Bruce M. Rowe and Diane P. Levine. 2015. *A Concise Introduction to Linguistics*. Routledge.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

Barbara Scalvini and Iben Nyholm Debess. 2024. Evaluating the potential of language-family-specific generative models for low-resource data augmentation: A Faroese case study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6496–6503, Torino, Italia. ELRA and ICCL.

Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 724–731, Dublin, Ireland. Association for Computational Linguistics.

Marco Antonio Sobrevilla Cabezudo, Marcio Lima Inacio, and Thiago Alexandre Salgueiro Pardo. 2024. Investigating paraphrase generation as a data augmentation strategy for low-resource AMR-to-text generation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 663–675, Tokyo, Japan. Association for Computational Linguistics.

Hongkai Sun. 2015. 541language policy of china's minority languages. In *The Oxford Handbook of Chinese Linguistics*. Oxford University Press.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma

Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Pe-

ter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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

Jörg Tiedemann and Preslav Nakov. 2013. Analyzing the use of character-level translation with sparse and noisy datasets. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 676–684, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Atnafu Tonja, Fazlourrahman Balouchzahi, Sabur Butt, Olga Kolesnikova, Hector Ceballos, Alexander Gelbukh, and Thamar Solorio. 2024. NLP progress in indigenous Latin American 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 6972–6987, Mexico City, Mexico. Association for Computational Linguistics.

Nicolas Tournadre. 2014. The tibetic languages and their classification. In Thomas Owen-Smith and Nathan Hill, editors, *Trans-Himalayan Linguistics*, pages 105–130. De Gruyter Mouton, Berlin, Boston.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Rekha Vijayvergia, Bharti Nathani, Nisheeth Joshi, and Rekha Jain. 2023. A survey on various approaches used in named entity recognition for indian languages. In *Proceedings of the 4th International Conference on Information Management & Machine Intelligence*, ICIMMI '22, New York, NY, USA. Association for Computing Machinery.

Minggang Wan and Shanxin Zhang. 2007. Chapter 7. research and practice of tibetan–chinese bilingual education. In Anwei Feng, editor, *Bilingual Education in China*, pages 127–144. Multilingual Matters, Bristol, Blue Ridge Summit.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lisa Woller, Viktor Hangya, and Alexander Fraser. 2021. Do not neglect related languages: The case of low-resource Occitan cross-lingual word embeddings. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 41–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.

Jinyi Zhang, Ke Su, Haowei Li, Jiannan Mao, Ye Tian, Feng Wen, Chong Guo, and Tadahiro Matsumoto. 2024. Neural machine translation for low-resource languages from a chinese-centric perspective: A survey. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(6).

Suxiang Zhang, Ying Qin, Juan Wen, and Xiaojie Wang. 2006. Word segmentation and named entity recognition for SIGHAN bakeoff3. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 158–161, Sydney, Australia. Association for Computational Linguistics.

A Sino-Tibetan Language Distribution

Figure 4 shows the distribution and classifications of Sino-Tibetan languages.

B Language Names Used

We list all the languages and the names that we used for query in Table 4 below:

C Conference Names

We include the full names of the conferences that appeared in our analysis:

1. **ACL**: Annual Meeting of the Association for Computational Linguistics
2. **IJCNLP**: International Joint Conference on Natural Language Processing
3. **LREC**: International Conference on Language Resources and Evaluation
4. **EMNLP**: Conference on Empirical Methods in Natural Language Processing
5. **CCL**: Chinese National Conference on Computational Linguistics
6. **PACLIC**: Pacific Asia Conference on Language, Information and Computation
7. **COLING**: International Conference on Computational Linguistics
8. **ROCLING**: Conference on Computational Linguistics and Speech Processing
9. **IJCLCLP**: International Journal of Computational Linguistics and Chinese Language Processing

D Method for Mini Experiment

We modified the zero-shot vanilla method proposed in Xie et al. (2023) by adding the example given in the paper. The prompt is as follows:

```
Given entity label set: {location,
    organization, person, miscellaneous}
Based on the given entity label set,
    please recognize the named entities
    in the given text.
Example:
Text: Could Tony Blair be in line for a
    gold medal?
Answer: {'Tony Blair': 'person'}

Text: <sentence>
Answer:
```

where `<sentence>` is a space-separated sentence constructed by concatenating words of the sentence in the dataset. We used the default setting of *gpt-3.5-turbo* and passed the prompt using the ChatGPT API[10] to generate an answer. We then parsed the answer given by ChatGPT, compared it to the gold labels and calculated the micro F_1 score.

[10]https://platform.openai.com/docs/api-reference/introduction

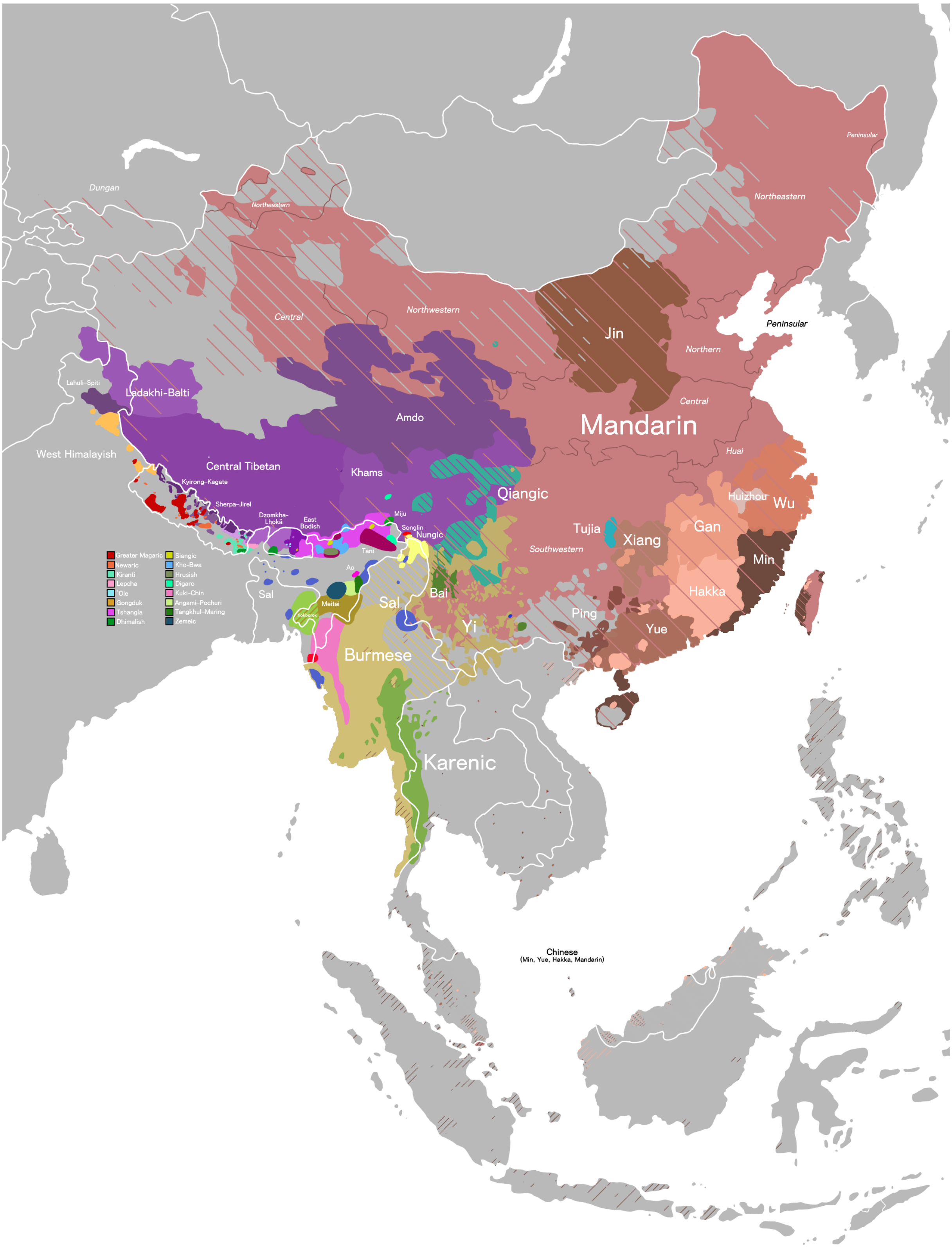

Figure 4: The distribution of Sino-Tibetan languages. Attribution: By GalaxMaps - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=115399902

Language	Query Terms
Akha	Akha
Amdo	Amdo, Amdo Tibetan
Angami	Angami
Ao	Ao Naga, Ao language, Ao
Arakanese (Marma)	Marma, Arakanese, Marma Arakanese, Rakhine
Bai	Bai language
Balti	Balti
Bodo	Bodo, Boro
Bokar	Adi, Bokar, Milang
Burmese	Burmese
Cantonese	Yue Chinese, Cantonese
Chin (Tiddim)	Tedim, Tedim Chin, Tiddim, Tiddim Chin
Dimasa	Dimasa
Fuzhou	Fuzhou, Foochow, Hokchew, Hok-chiu, Fuzhounese, Min Dong Chinese
Garo	Garo
Gurung	Gurung, Tamu Kyi, Tamu Bhasa
Hakka	Hakka, Hakka Chinese
Hani	Hani
Hyow	Hyow, Asho, Asho Chin
Jingpho	Jingpho, Jingpo, Jinghpaw, Kachin
Karen (Pwo)	Pwo, Pwo Karen
Karen (Sgaw)	S'gaw, S'gaw Karen, S'gaw K'Nyaw
Kham (Tibetan) (Nangchen)	Khams Tibetan, Khams
Kokborok	Kokborok, Kok Borok
Lahu	Lahu
Lai	Hakha, Hakha Chin, Laiholh
Limbu	Limbu
Lisu	Lisu, Lisu language
Lotha	Lotha, Lotha Naga
Magar	Eastern Magar
Magar (Syangja)	Syangja, Magar Syangja, Syangja Magar, Western Magar
Maru	Lhao Vo, Maru, Lhaovo
Meithei	Meithei, Meitei
Mikir	Mikir, Karbi
Min (Southern)	Min Nan, Min Nan Chinese, Hokkien, Southern Min, Banlam
Mising	Mising
Mizo	Mizo
Naga (Zeme)	Zeme
Naxi	Naxi, Nakhi, Nasi, Lomi, Moso, Mo-su
Newari (Kathmandu)	Newar, Newari
Nuosu	Nuosu, Nosu, Northern Yi, Liangshan Yi, Sichuan Yi
Nyishi	Nyishi, Nishi, Nisi, Nishang, Nissi, Nyising, Leil, Aya, Akang, Bangni-Bangru, Solung
Sherpa	Sherpa
Tamang (Eastern)	Tamang
Thadou	Thadou, Thado Chin
Tibetan (Modern Literary)	Tibetan
Tshangla	Tshangla
Wu	Wu Chinese, Suzhounese, Shanghainese
Yi (Wuding-Luquan)	Luquan Yi, Wuding Yi, Nasu

Table 4: The complete list of all languages covered in our study and the query terms that we used when searching through the corpus