The Zeno's Paradox of 'Low-Resource' Languages

Hellina Hailu Nigatu[1, *] **Atnafu Lambebo Tonja**[2,3,]
Benjamin Rosman[3,4, †] **Thamar Solorio**[2,5 †] **Monojit Choudhury**[2,†]
Corresponding author: hellina_nigatu@berkeley.edu

[1] UC Berkeley, USA, [2] MBZUAI, UAE, [3] Lelapa AI, South Africa
[4] RAIL Lab - University of the Witwatersrand, South Africa, [5] University of Houston, Houston, USA

Abstract

The disparity in the languages commonly studied in Natural Language Processing (NLP) is typically reflected by referring to languages as low vs high-resourced. However, there is limited consensus on what exactly qualifies as a 'low-resource language.' To understand how NLP papers define and study 'low resource' languages, we qualitatively analyzed 150 papers from the ACL Anthology and popular speech-processing conferences that mention the keyword 'low-resource.' Based on our analysis, we show how several interacting axes contribute to 'low-resourcedness' of a language and why that makes it difficult to track progress for each individual language. We hope our work (1) elicits explicit definitions of the terminology when it is used in papers and (2) provides grounding for the different axes to consider when connoting a language as low-resource.

1 Introduction

> If the fleet-footed Achilles and a slow-moving tortoise are in a race, Achilles will never catch the tortoise if the tortoise has a head start. Regardless of how fast Achilles runs, he first has to reach a point the tortoise already passed, by which point the tortoise will have moved ahead. –Zeno's Achilles Paradox [1]

The majority of research in the NLP community has focused on only a handful of the world's languages (Joshi et al., 2020; Bird, 2022). Particularly, languages spoken by communities in the Global South have largely been neglected (Nekoto et al., 2020; Schwartz, 2022). Languages understudied by the NLP community are usually referred to as 'low-resource', while those well-studied are referred to as 'high-resource.' This framing of high vs low-resource languages resembles Zeno's Achilles paradox: 'high-resourced languages' are the tortoise, that have been given a head start in the research community and continue to receive much of the attention, and 'low-resource languages' are Achilles. In reality, Achilles can always outrun the tortoise[2]. However, the face value interpretation of the paradox can serve as an analogy for how the current trajectory of the NLP research community to include majority of the worlds languages in the path already forged for 'high-resourced' languages leaves 'low-resource languages' constantly trying to catch up to a goalpost that is always moving.

The disparity in research and performance of language technologies across languages can be a double-edged sword. On the one hand, understudied and underserved languages may be at a higher risk of language loss and have speakers exposed to direct downstream harm due to failures of language technologies (Nigatu and Raji, 2024; Choudhury, 2023). On the other hand, the drive to include these languages in research without proper consideration of community needs (1) may lead to aggressive–and at times exploitative–data collection and (2) result in technologies that do not meet the needs of the communities who speak those languages (Diddee et al., 2022; Le Ferrand et al., 2022a; Dearden and Tucker, 2021).

Recently, we have seen efforts to increase the representation of 'low-resource languages' in NLP research (e.g. NLLB, 2024; Adelani et al., 2022). Yet, the exact definition of the term 'low-resource' remains elusive[3]. A common criterion to connote languages as 'low' vs 'high' resourced is data. However, using data as the only criterion oversim-

*Work done while this author was at MBZUAI.
†These authors provided equal advice and supervision.
[1]https://www.britannica.com/topic/Achilles-paradox
[2]https://ibmathsresources.com/2018/11/30/zenos-paradox-achilles-and-the-tortoise-2/
[3]'Under-resource' is a term used interchangeably–and perhaps equally as ambiguously–with 'low-resource.' For brevity, we mainly use the phrase 'low-resource' in this paper.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17753–17774
November 12-16, 2024

plifies the context of the language itself. Languages dubbed as 'low-resource' may vary depending on factors like their number of speakers, non-digital archives, or language experts (Kuhn, 2024).

The lack of consensus in what qualifies a language as 'low-resource' makes it challenging to **(1)** track progress in research and development for 'low-resource languages' in general, **(2)** determine what interventions are effected towards a language, **(3)** pinpoint when a language stops being 'low-resource', and **(4)** discern if technologies built for these languages truly address the needs of the communities who speak them or if they are built simply on the premise that the same technology exists for a 'higher resourced language.'

In this work, we survey papers that study languages coined as 'low-resource'. We qualitatively analyzed 150 papers that include the keywords 'low-resource' and 'under-resource.' We used qualitative methods to unravel (1) how such papers define the term 'low-resource' or 'under-resource', (2) what languages are studied as 'low-resource', and (3) what criteria is used to classify a language as 'low-resource.'

Our analysis reveals four separate but interacting aspects of 'resourcedness' that are used to connote a language as 'low-resource' (see Section 3 & Section 4). In Section 5, we use real-world examples to demonstrate how each of the aspects interact and how those interactions impact what interventions are designed and implemented for a language. Finally, we use our analysis to ground recommendations for different stakeholders (see Section 6).

2 Methodology

Data We collected data for papers published at *CL venues[4] from the ACL Anthology[5] and at the following Speech Processing conferences: INTERSPEECH and International Conference on Acoustics, Speech, and Signal Processing (ICASSP) using the Semantic Scholar (Kinney et al., 2023) API . We used a keyword search to identify papers that include the terms 'low-resource' or 'under-resource' in their titles or abstracts. Our final corpus included 868 unique papers.

Qualitative Analysis In the initial stage of our analysis, we found that the term 'low-resource' is used to refer to three broad categories: (1) tasks and domains where there is a lack of labeled data, (2) 'simulated low-resource' settings via methods like under-sampling , (3) 'low-resource languages' defined based on diverse criteria. Table 1 summarizes this finding. For our qualitative analysis, we exclusively focused on the third category, i.e., papers that study 'low-resource languages' as our interest is in understanding how a language is labeled as low-resource. We also found papers that tried both sampling higher-resourced languages and using actual, low-resourced languages (e.g. Zevallos and Bel, 2023b). We include those in our analysis as they study a 'low-resourced language' in addition to a simulated setting. We manually labeled 541 papers to identify those that explicitly work on non-simulated low-resource languages and randomly sampled 150 papers for qualitative analysis. Our sampling strategy was independent of any parameter such as publication year; the time span for the 150 papers was 2017-2023. We conducted our analysis by reading each paper and annotating how the term 'low-resource' or 'under-resource' is defined, what languages are studied in the paper, and any additional challenges mentioned in the paper in relation to the languages of study being 'low-resource.' We used inductive thematic analysis (Braun and Clarke, 2006) and discussed the themes that emerged from our analysis in frequent meetings to synthesize overarching themes. In the following section, we present the results of our analysis along with illustrative quotes.

Category	Description	Examples	%
Tasks and Domains	tasks and domains where there is limited labeled data	Sun et al. (2022a); Bajaj et al. (2021)	27.27
Simulated	using techniques like under-sampling to simulate low-resource settings	Zevallos and Bel (2023b); Dehouck and Gómez-Rodríguez (2020)	12.27
Languages	languages categorized based on factors like data or number of speakers	Coto-Solano (2022); Ponti et al. (2021)	65.04

Table 1: **Three categories of papers returned for the keyword search for 'low-resource.'** Note that the percentages do not add up to 100 because some papers fall into more than one category. For instance, Mager et al. (2020) study both simulated and actual low-resource languages.

[4]We focused on the top 6 venues based on Google Scholar metrics for computational linguistics(https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computationallinguistics)

[5]https://github.com/acl-org/acl-anthology

3 What is a 'low-resource' language?

In this section, we present the overarching aspects we found from our thematic analysis. It is first important to note the different styles papers use when defining the term 'low-resource':

> *"Languages facing this lack of large amount of data are called low-resourced, and all linguistic varieties in Mexico are struggling with this situation."* – Sierra Martínez et al. (2020)

> *"Under-resourced, under-studied and endangered or small languages yield problems for automatic processing and exploiting because of the small amount of available data as well as the missing or sparse description of the languages."* –Ferger (2020)

> *"It frames these as "low resource languages," lacking the text, speech and lexical resources that are needed for creating speech and language technologies (Krauwer, 2003)".* –Lane and Bird (2021a)

In the quotes shown above we see that Sierra Martínez et al. (2020) explicitly define the term, Ferger (2020) describes challenges of working with low-resource and Lane and Bird (2021a) define the term and provide citations from prior work. If a paper uses prior work without explicitly stating its definition, we rely on the definition of the cited work. In cases where there are no explicit definitions, we rely on the challenges mentioned by the paper to categorize how the paper decides if a language is 'low-resource.'

We found that definitions for the term 'low-resource' borrow from four aspects: (1) **Socio-political** aspects relating to financial and historical constraints, (2) **Resources**, both human and digital, (3) **Artifacts** such as linguistic knowledge, data, and technological infrastructure, and (4) **Agency** of community members in what technology is built for their languages. We summarize these four aspects in Figure 1 and dive into detail about each aspect in the following subsections.

3.1 Socio-Political

Some papers call out structural issues pertaining to societal, economic, and political forces. We found papers that reflect on low-resourcedness due to financial and economic constraints to curating data (e.g. Coto-Solano, 2022; Pathak et al., 2022) and limited use of such languages in mainstream media, government, and education (e.g. Mehta et al., 2020). For example:

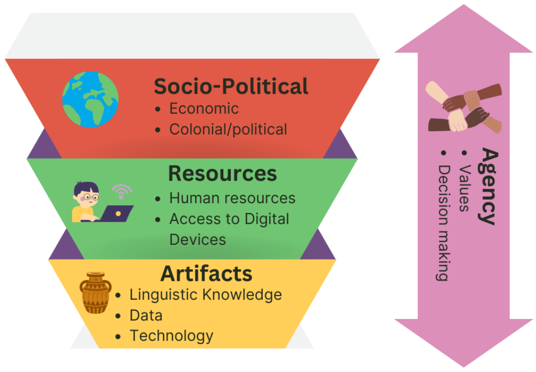

Figure 1: **Four overarching aspects that contribute to a language being classified as low-resource.** Socio-political aspects are at the top, influencing both the availability of resources and the creation of artifacts. Community agency is a common thread in all the other three aspects.

> *"In many of these communities, languages like English and Spanish have displaced the Indigenous languages in domains such as technology and chatting, and so the available data is curtailed."*– Feldman and Coto-Solano (2020)

> *"However, these languages are not represented in education, government, public services, and media, and therefore, they show high levels of endangerment."*– Sierra Martínez et al. (2020)

3.2 Resource

The second aspect discussed by papers is the availability of and access to human and digital resources[6].

Human Resources We found three types of human resources mentioned in papers in relation to low-resource languages: (1) native speakers (e.g. Feldman and Coto-Solano, 2020; Leong et al., 2022), (2) linguistic experts (e.g. Pathak et al., 2022), and (3) NLP researchers (e.g. Yimam et al.,

[6] Note that in the context of this work, data is an artifact curated for NLP purposes and so is not referred to as a resource in this category.

2020). With regards to native speakers, while some low-resource languages are described as having a limited number of native speakers, others are described as still being low-resourced despite a large number of native speakers. For instance:

> *"Quechua, a low-resource language from South America, is a language spoken by millions but, despite several efforts in the past, still lacks the resources necessary to build high-performance computational systems."*–Melgarejo et al. (2022)

> *"However, low-resource languages such as Amharic have received less attention due to several reasons such as lack of well-annotated datasets, unavailability of computing resources, and fewer or no expert researchers in the area."*–Yimam et al. (2020)

Access to Digital Devices and Platforms Lack of access to digital devices–and by extension, the digital presence of communities–is another reason mentioned in relation to 'low-resource' languages (e.g. Bamutura et al., 2020; Nzeyimana and Niyongabo Rubungo, 2022). Mainly, this reason is tied to the lack of available digital data for languages that fit the mainstream way of training models. Papers state that 'low-resource' languages are not available in formats suitable for crawls and scraping (e.g. Feldman and Coto-Solano, 2020).

> *"The included low-resource languages are also very limited because they are mainly sourced from Wikipedia articles, where languages with few articles like Kinyarwanda are often left behind."* – Nzeyimana and Niyongabo Rubungo (2022)

> *"In addition to this, many Indigenous communities have chronic digital inequalities, which makes it difficult to generate crowd-sourcing campaigns for those languages. Finally, in many cases, the data that is most valuable to speakers of the language is collected from elders and knowledge keepers, but those elders might be the people who have the least access to technological means of communication."* –Feldman and Coto-Solano (2020)

3.3 Artifacts

The third aspect of resourcedness is tied to the production and accessibility of artifacts: linguistic knowledge, data, and technology.

Linguistic Features and Descriptions Papers state how there are limited available linguistic descriptions for 'low-resource' languages (e.g. Ferger, 2020; Sikasote and Anastasopoulos, 2022). Often, linguistic features–such as morphological complexity and typology–are used as reasons why it is difficult to blindly adopt methods that work for high-resource languages, even in cases where there is an equal number of training data (e.g. de Lhoneux et al., 2022). Standardization–or lack thereof–is another feature mentioned in relation to 'low-resourcedness' of languages. Both linguistic features and lack of standardization are mentioned as reasons for data sparsity. For example:

> *"Due to differences in language typology, it is not necessarily as simple as looking only at number of lines of training data.[...] For example, Inuktitut is known to be highly morphologically complex, resulting in many words (defined as space/punctuation separated) that appear just once or only a few times, even in such a large corpus."*–Knowles and Littell (2022)

> *"Not only is data scarce, but it might lack standardization, making the dataset more sparse than it would be for languages with standardized orthographies and numerous speakers."* –Coto-Solano (2022)

Data With regards to data, the classification of a language as low-resource could be based on labeled or annotated data (e.g. Ponti et al., 2021), unlabeled data (e.g. ImaniGooghari et al., 2022), or benchmark data (e.g. Reid et al., 2021). Some papers focus their definitions on the quality of data (e.g. Maillard et al., 2023; Ramnath et al., 2021), stating that low-resource language data is usually noisy. Other papers quantify the amount of data (e.g. Biswas et al., 2020). We also observed a subset of papers that use a predefined cutoff for the amount of data: for instance, Ramachandran and de Melo (2020) state they *"...picked six languages that had around 10K or fewer verses available."* Some papers would quantify the amount of data in relation to a popular trend in the field:

"Only some of the 22 scheduled Indian languages, which are a subset of the numerous languages spoken and written in India, have enough resources for training a deep learning model." –Saurav et al. (2020)

Technology Exclusion from technological advances for the languages of study is another aspect mentioned in relation to low-resource languages. This ranges from the lack of basic computational tools–such as text pre-processing tools (e.g. Niyongabo et al., 2020) –to exclusion from pre-trained language models (e.g. Leong et al., 2022; Pfeiffer et al., 2020). There were also mentions of lack of compute resources (e.g. Yimam et al., 2020).

"Handling utterances with non-Kanien'kéha characters would have required grapheme-to-phoneme prediction capable of dealing with multilingual text and code-switching, which we did not have available." –Pine et al. (2022a)

"In total, we can discern four categories in our language set: 1) high-resource languages and 2) low-resource languages covered by the pretrained SOTA multilingual models (i.e., by mBERT and XLM-R); as well as 3) low-resource languages and 4) truly low-resource languages not covered by the multilingual models"–Pfeiffer et al. (2020)

3.4 Agency

Transcending all the other aspects is community agency and the role it plays in what and by whom language technologies are built. Coto-Solano (2022) state how even in cases where communities are willing to provide data, financial constraints prevent them from doing so. Le Ferrand et al. (2022a) emphasize building language tools detached from community practices leads to technologies with minimal utility to the communities. This detachment from community practices is also stated as a reason for minimal studies in these languages:

"Although Assamese has a very old and rich literary history, technology development in NLP is still in a nascent stage." –Pathak et al. (2022)

When communities are actively engaged, we observe their values embedded in the production of technology, regardless of the outcome of the research project:

"While a total of 24 hours of audio were recorded, members of the Kanien'kéha-speaking community told us it would be inappropriate to use the voices of speakers who had passed away, leaving only recordings of Satewas's voice. [...] The resulting speech corpus comprised 3.46 hours of speech." –Pine et al. (2022b)

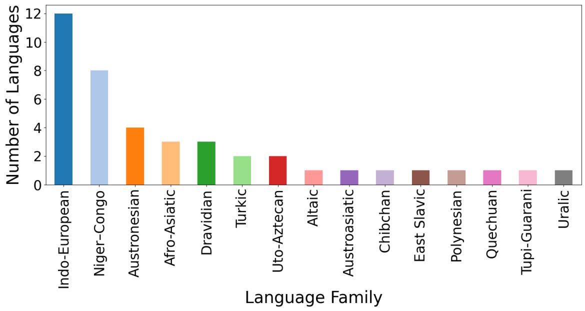

Figure 2: Number of languages included in the studies per language family.

4 What Languages are Studied as 'Low-Resource'?

Languages may be studied in multilingual contexts, i.e. included alongside other languages (e.g. Adelani et al., 2022; Goyal et al., 2021) or in monolingual contexts (e.g. Yimam et al., 2020; Pathak et al., 2022). Papers had varying depths of descriptions for the languages they studied, with papers working on fewer languages having more in-depth descriptions. For instance, Mehta et al. (2020), which exclusively work on the Gondi language, has a dedicated section on the historical, political, and linguistic context of the Gondi language and its community. On the other hand, Goyal et al. (2021), which works on 101 languages, has one table with all the languages, their ISO codes, language families, writing scripts, and the amount of available data.

In Figure 2, we show the number of languages and language families studied in our samples, where papers explicitly mention them as low-resource. We observe a diverse set of language families, with Indo-European languages having the highest number of languages studied in our samples, followed by Niger-Congo and Austronesian. In Appendix C, we detail the top 20 most frequently studied languages in our sample.

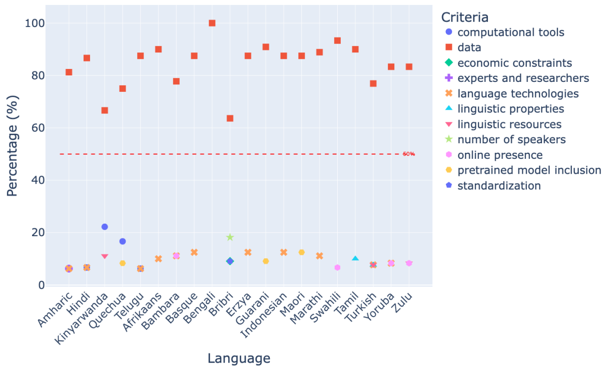

Figure 3: Criteria distribution used in the top-20 languages to categorize languages.

The graph in Figure 3 shows the distributions of the various criteria used for categorizing a language as 'low-resource' in the top 20 languages studied. While data is the most commonly used criterion across many papers and languages, other factors, such as lack of computational tools, limited number of native speakers, etc, are also used (see Section 3). Even with papers that use data as a criterion, we observe different qualifications for *what type* of data a language may lack to qualify as a 'low-resource' language. In Figure 6, we further break down the criterion of data. We observe that lack of labeled data is the most commonly used criterion in our sample at 39.8%. We also observe the lack of digitized text (1.7%) and online-available data (6.9%) as criteria to connote a language as low-resource.

5 Why does it matter?

In the previous section, we describe four overarching aspects that determine if a language is 'low-resource': socio-political aspects, human and digital resources, artifacts, and agency of community members. In Figure 4, we present language profiles for 6 languages. We choose the six languages from the bottom three classes in Joshi et al. (2020): 'The Left Behinds' with limited labeled and unlabeled data, 'The Scraping-Bys' with some amount of unlabeled data, and 'The Hopefuls' with some labeled data. We use literature about these languages and their communities to demonstrate why it matters that we are specific in the terminology we use.

Languages in the same class of data availability might differ in other aspects. From 'The Left Behinds', we present profiles for Numma-guhooni[7] and Warlpiri. Numma-guhooni is spoken in Kenya where the official Federal languages are Kiswahili[8] and English. Warlpiri is spoken by the Warlpiri people of Australia, where the most dominant language is English. While both languages fall into the same class, the number of speakers for Warlpiri is 4 times that of Numma-guhooni. Ethnologue classifies Warlpiri as a *stable* language, while Numma-guhooni is *endangered*. In terms of digital resource availability, Ethnologue classifies Numma-guhooni as *still* meaning, there is no sign of digital support for the language, while Warlpiri is labeled *emerging* with some digital content available. Warlpiri also has some NLP tools available, for instance, KirrKirr is a dictionary visualization tool for the Warlpiri language (Manning et al., 2001).

From 'The Scraping-Bys', we look at Cherokee and Kalaallisut. Cherokee, spoken by around 2,000 out of the 300,000 Cherokee people of the Cherokee Nation in the United States of America, is labeled as *endangered* by Ethnologue. On the other hand, Kalaallisut, which is spoken by about 50,000 people and is the official Federal language of Greenland, is labeled as *institutional* by Ethnoluge. However, Cherokee has a higher ranking for digital language support, dubbed *vital* while Kalaallisut is *ascending*.

For 'The Hopefuls', we look at isiZulu and Konkani. We observe the two languages are somewhat similar in terms of human and digital resources, with both being *institutional* in vitality and *vital* in digital access. However, we see the languages vary by their number of speakers with isiZulu having about 6 times the number of speakers as Konkani. Additionally, isiZulu is the most common language spoken as a first language in South Africa, while Konkani has shown a decline in number of speakers, with speakers outside of its primary province declaring other, dominant languages as their native language (Rajan et al., 2020). Both languages have NLP tools available for tasks like machine translation and speech processing as well as pre-processing tools.

Overall, we observe that within a given class based on data availability, there are drastic differences in what other resources are available for a

[7]While this language is refereed to with another name in the literature, there is evidence that the word is derogatory and so we exclusively use the name native speakers use (Stiles, 1982).

[8]also known as Swahili in English speaking contexts.

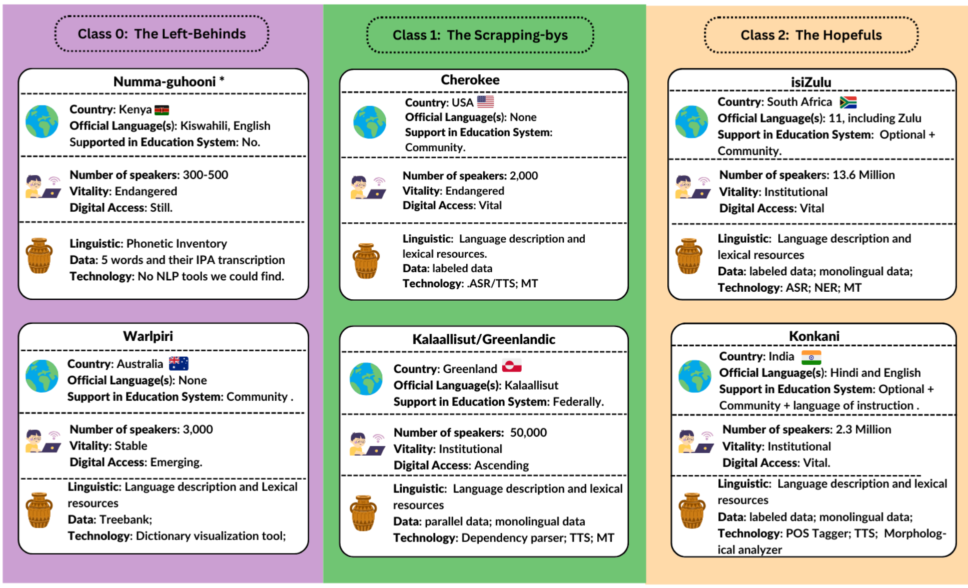

Figure 4: **Language profiles for six languages across three classes based on data availability.** The first row in each profile deals with socio-political issues, the second row resources, and the last row with artifacts (see Figure 1). We observe drastic differences between languages of the same class. See Appendix A for details on the labels.

language. We observe that the variance decreases as we move up the classes, which can partially be explained by the stark 88.38% of the world's languages belonging to 'The Left-Behinds', compared to 5.49% in 'The Scraping-Bys' and 0.36% in 'The Hopefuls' (Joshi et al., 2020). However, as we demonstrate, the realities of each of the languages within each class are very different.

The different aspects that determine 'low-resourcedness' have causal links. The four aspects we discuss in Section 3 interact with each other in constraining what is available. **Socio-political** issues constrain what **Resources** are available for a given language, which in turn impact what **Artifacts** are produced for that language. For instance, while there are no official languages in the USA or Australia, federal policies in the US up to 1948 forced Indigenous children to assimilate into Western culture, punishing students for speaking their languages (Wakeman, 2021). Similarly, colonization destroyed several languages of Indigenous populations in Australia (Laura Stocker and Rooney, 2016). As a result, both Cherokee and Warlpiri, along with the numerous other Indigenous languages of the Americas, Australia, and Canada are endangered, i.e lack **human resources**.

Assimilation is not limited to the languages of the colonizer. Post-independence from colonial rule of Britain, Kenya adopted the educational and language policies of Britain, with English declared the official language in formal sectors and Kiswahili the national language of the country. As a result, the majority of data available in **digital and electronic media** as well as in public settings are in English or Kiswahili (Barasa, 2023). Hence, speakers of languages like Numma-guhooni are largely assimilated with larger ethnic groups and Kiswahili is predominantly spoken and learned by the new generation (Tosco, 1992). While in 2010, the Kenya constitution shifted towards centering the preservation of native languages, there were not enough funds allocated to carry this through (Barasa, 2023). Though at a different scale, this is similar to the case of Konkani, which is in 'The Hopefuls' class, losing native speakers to more dominant local languages (Rajan et al., 2020).

Constraints of human and digital resources restrict the creation of **artifacts** for languages. As discussed in Section 3.2, the minimal digital presence results in limited **available data**, especially at the scale needed for training SOTA **models.** Links among the different aspects are not necessarily linear; socio-political issues also directly constrain what languages are taught in schools, impacting **linguistic knowledge** produced for a language. Ad-

Aspect	Sub-Division	Terminology	Definition
Socio-political	Economic	low-affluence (Hammarström, 2009)	based on Gross Language Product (GLP) (product of the number of native speakers in any country and the country's per capita Gross National Product.)
	Political	politically-disadvantaged	languages not used in mainstream media and governmental communications due to political forces
Resources	Native Speakers*	extinct; critically endangered; severely endangered; definitively endangered; unsafe; safe (Brenzinger et al., 2003)	6 point scale based on number of speakers of the language
	Online Presence	Low-Web Resource(Patil et al., 2022)	limited online corpus or web presence
	Language experts	expert-constrained	limited number of linguistic experts or researchers
Artifacts	Linguistic Knowledge*	oral languages; non-native orthography; native orthography	based on the availability and type of orthography a language has.
		undocumented; inadequate; fragmentory; fair; good; superlative (Brenzinger et al., 2003)	6 point scale based on the amount and quality of documentation available for a language.
	Data*	Class 0; Class 1; Class 2; Class 3; Class 4; Class 5 (Joshi et al., 2020)	6 classes based on the availability of labeled and unlabeled data
	Technology*	Still; Emerging; Ascending; Vital; Thriving (Simons et al., 2022)	5-level classification based on digital language support available in a given language.

Table 2: **Suggestions for explicit terminology addressing three aspects we identified through our analysis.** We provide citations for terminology taken from prior work. (*) indicate the terminology are part of a scale and all labels in the scale are listed.

ditionally, prior work demonstrates the Western-dominated **researcher** landscape in NLP and how it ties to coloniality (Held et al., 2023). With the limited number of speakers for a given language, the number of NLP researchers who are also native speakers of the language is largely constrained, which is further confounded by the limited financial resources available to researchers from such communities. As a result, having **agency** in what tools are designed for a language becomes challenging.

Knowing which aspect a language is lacking in allows for targeted interventions. One of the main factors that determine the survival of a language is *inter-generational transmission* (Brenzinger et al., 2003). For instance, while Cherokee and Kalaallisut are both in the same class, Cherokee is *endangered* while *Kalaallisut* is institutional. Hence, interventions–both in socio-political and artifact aspects–are best targeted toward reviving and preserving the Cherokee language. On the other hand, digital access for Kalaalisut is *ascending*, hence there might be more efforts towards increasing the availability of digital data. Since Kalaallisut is institutional, financial resources for preserving and growing the language are available at a federal level. Additionally, it is used as the language of instruction in the education system of the country, aiding in the inter-generational transfer of the language. Across classes, we observe similarities in Numma-guhooni and Konkani, of native speakers assimilating to other dominant but local languages. Hence, interventions for these languages may be more effective in language learning apps that focus on learning the less-dominant language and translation systems between dominant local languages and the target language.

Communities are actively resisting exploitation and sustaining their languages; our tools should support them. Despite the several layers of constraints, it is important to note that communities are not in idle state of deficit. Across classes, we observe a similarity between Warlpiri and Cherokee, in that there are community-based initiatives to preserve and grow the languages (e.g. the Warlpiri Education and Training Trust (WETT)[9] and the Cherokee Immersion School[10]). By centering community values in our designs and research, we can collectively forge new paths for each language, conditioned on its unique circumstances.

[9]https://www.clc.org.au/wett/
[10]https://www.cwyschools.org/

6 What can we do?

Using specific terminology or having explicit definitions allows us to measure progress more precisely. The specific *resource* a language is deemed 'low' in directly impacts what interventions are effected towards it. For instance, programs aimed at increasing language representation in Human Language Technologies (HLTs) have several selection criteria (Cieri et al., 2016). Such programs use different terminologies and definitions, where "each term encodes differences in traditions, goals, and approaches" (Simpson et al., 2008). As a result, what languages are included and served by such programs differ, even if languages have the same amount of data.

While Cherokee is tagged as having *vital* digital resources, it is also an *endangered* language. Collecting more data in the language from the limited number of speakers or including it in Large Language Models may not exactly alleviate its low-resourcedness. We argue for more explicit declarations of *which* aspects of resources are being referred to when the term low-resource is used. In Table 2, we give recommendations for terminologies based on prior work and our findings. There are also several taxonomies and classes based on data (e.g Joshi et al., 2020), language vitality (e.g Brenzinger et al., 2003), and digital support (e.g Simons et al., 2022).

Recommendations for stakeholders: Based on our findings, we give recommendations for different stakeholders involved in the effort to increase language representation in NLP research. Individual researchers can (1) engage with community members and speakers of the languages they work on, (2) articulate how their work is limited in relation to the characteristics of the languages they work on, and (3) be explicit about what criteria they use to denote a language as 'low-resource.' Community members can also form grassroots organizations such as Masakhane[11], which allow researchers who speak diverse languages to build language technologies together and learn from each other's experiences. Additionally, such organizations can prioritize engaging with native speakers who may not be in the NLP research field, allowing for diverse perspectives when deciding what tools should be built for what language. Workshops such as AmericasNLP[12] and AfricaNLP[13] continue to serve as spaces for fostering research and collaboration for languages that are mostly ignored in mainstream NLP research. However, main (*)CL conferences can increase the representation of these languages by (1) offering alternative tracks for papers, (2) easing the cost of attendance and registration for researchers from these communities, and (3) diversifying conference venues. Academic institutions can aid researchers who speak these languages by promoting interdisciplinary collaboration and partner with local and international organizations to document and preserve marginalized languages. Industry players interested in language diversity of their products can play a role by offering financial and technical support; for example, subsidizing resources for communities working on low-resource languages. Companies could also prioritize making their products accessible to the communities (e.g. Üstün et al., 2024). Government bodies can play a role in preserving languages through policies, funding, and digital inclusion. Funding agencies can support language diversity and enforce building technologies that are relevant to the specific linguistic community by setting research priorities and prioritizing grants to underrepresented researchers.

7 Conclusion

In this paper, we present 4 aspects of 'resourcedness' used to classify a language as 'low-resource' based on a qualitative survey of 150 papers. Based on our analysis, we give recommendations for terminology that explicitly calls out which resource we are referring to when we say a language is 'low-resource.' A language may lack in several aspects, making the use of individual terminology difficult– e.g. in multilingual settings. However, the difficulty does not absolve us from the responsibility to provide detailed documentation. At the very least, clear statements on what exactly is meant by low-resource when referring to a language would allow us to more clearly articulate the problems a particular technology resolves for a particular language.

8 Limitations

As a qualitative study, our paper does not give the definitions of the term from all the papers in all the

[11]https://www.masakhane.io/

[12]https://github.com/AmericasNLP

[13]https://africanlp.masakhane.io/

venues we searched. We also do not make quantitative claims. Instead, we focus on a nuanced analysis of how our sample papers describe the phenomenon and provide direct quotes from papers we analyzed as evidence. While it was not practical for us to conduct qualitative analysis on more than the papers in our sample, future work could use automated methods and conduct a quantitative analysis. Similarly, our analysis of what languages are studied is limited to the papers in our sample. This could also be supplemented with automated extraction at scale. Additionally, while we could not perform a longitudinal analysis with our sample size of 150 papers, future work could explore such a study to understand how the use of the term 'low-resource' evolved over time.

References

Ife Adebara, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2022. Linguistically-motivated Yorùbá-English machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5066–5075, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Wafia Adouane, Samia Touileb, and Jean-Philippe Bernardy. 2020. Identifying sentiments in Algerian code-switched user-generated comments. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2698–2705, Marseille, France. European Language Resources Association.

Kabir Ahuja, Sunayana Sitaram, Sandipan Dandapat, and Monojit Choudhury. 2022. On the calibration of massively multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4310–4323, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Karl Anderbeck. 2015. *Portraits of language vitality in the languages of Indonesia*, pages 19–47.

Andrei-Marius Avram, Darius Catrina, Dumitru-Clementin Cercel, Mihai Dascalu, Traian Rebedea, Vasile Pais, and Dan Tufis. 2022. Distilling the knowledge of Romanian BERTs using multiple teachers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 374–384, Marseille, France. European Language Resources Association.

Ahsaas Bajaj, Pavitra Dangati, Kalpesh Krishna, Pradhiksha Ashok Kumar, Rheeya Uppaal, Bradford Windsor, Eliot Brenner, Dominic Dotterrer, Rajarshi Das, and Andrew McCallum. 2021. Long Document Summarization in a Low Resource Setting using Pretrained Language Models. 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 71–80, Online. Association for Computational Linguistics.

David Bamutura, Peter Ljunglöf, and Peter Nebende. 2020. Towards computational resource grammars for Runyankore and rukiga. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2846–2854, Marseille, France. European Language Resources Association.

David Barasa. 2023. Language ideologies, policies and practices within the multilingual Kenyan context. *JLLCS*, 2(1):55–62.

M Saiful Bari, Tasnim Mohiuddin, and Shafiq Joty. 2021. UXLA: A robust unsupervised data augmentation framework for zero-resource cross-lingual NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages

1978–1992, Online. Association for Computational Linguistics.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

Martijn Bartelds and Martijn Wieling. 2022. Quantifying language variation acoustically with few resources. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3735–3741, Seattle, United States. Association for Computational Linguistics.

Vineet Bhat, Preethi Jyothi, and Pushpak Bhattacharyya. 2023. Adversarial training for low-resource disfluency correction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8112–8122, Toronto, Canada. Association for Computational Linguistics.

Steven Bird. 2022. Local Languages, Third Spaces, and other High-Resource Scenarios. *ACL Anthology*, pages 7817–7829.

Astik Biswas, Emre Yilmaz, Febe De Wet, Ewald Van der westhuizen, and Thomas Niesler. 2020. Semi-supervised development of ASR systems for multilingual code-switched speech in under-resourced languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3468–3474, Marseille, France. European Language Resources Association.

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*.

Matthias Brenzinger, Arienne M. Dwyer, Tjeerd de Graaf, Colette Grinevald, Michael Krauss, Osahito Miyaoka, Nicholas Ostler, Osamu Sakiyama, María E. Villalón, Akira Y. Yamamoto, and Ofelia Zepeda. 2003. UNESCO Ad Hoc Expert Group on Endangered Languages.

Jacqueline Brixey, David Sides, Timothy Vizthum, David Traum, and Khalil Iskarous. 2020. Exploring a Choctaw language corpus with word vectors and minimum distance length. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2746–2753, Marseille, France. European Language Resources Association.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No data to crawl? monolingual corpus creation from PDF files of truly low-resource languages in Peru. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Joan Byamugisha. 2022. Noun class disambiguation in Runyankore and related languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4350–4359, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Aditi Chaudhary, Antonios Anastasopoulos, Zaid Sheikh, and Graham Neubig. 2021. Reducing confusion in active learning for part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 9:1–16.

Nuo Chen, Linjun Shou, Ming Gong, Jian Pei, and Daxin Jiang. 2022. Bridging the gap between language models and cross-lingual sequence labeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1909–1923, Seattle, United States. Association for Computational Linguistics.

Monojit Choudhury. 2023. Generative AI has a language problem. *Nat. Hum. Behav.*, 7:1802–1803.

Chiamaka Chukwuneke, Ignatius Ezeani, Paul Rayson, and Mahmoud El-Haj. 2022. IgboBERT models: Building and training transformer models for the Igbo language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5114–5122, Marseille, France. European Language Resources Association.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection Criteria for Low Resource Language Programs. *ACL Anthology*, pages 4543–4549.

Rolando Coto-Solano. 2022. Evaluating word embeddings in extremely under-resourced languages: A case study in Bribri. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4455–4467, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. 2022. Development of automatic speech recognition for the documentation of Cook Islands Māori. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3872–3882, Marseille, France. European Language Resources Association.

Jeanne E. Daniel, Willie Brink, Ryan Eloff, and Charles Copley. 2019. Towards automating healthcare question answering in a noisy multilingual low-resource setting. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 948–953, Florence, Italy. Association for Computational Linguistics.

Miryam de Lhoneux, Sheng Zhang, and Anders Søgaard. 2022. Zero-shot dependency parsing with worst-case aware automated curriculum learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 578–587, Dublin, Ireland. Association for Computational Linguistics.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Andy Dearden and William D. Tucker. 2021. The ethical limits of bungee research in ICTD. In *2015 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–6. IEEE Press.

Arnab Debnath, Navid Rajabi, Fardina Fathmiul Alam, and Antonios Anastasopoulos. 2021. Towards more equitable question answering systems: How much more data do you need? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 621–629, Online. Association for Computational Linguistics.

Mathieu Dehouck and Carlos Gómez-Rodríguez. 2020. Data Augmentation via Subtree Swapping for Dependency Parsing of Low-Resource Languages. *ACL Anthology*, pages 3818–3830.

Prajit Dhar, Arianna Bisazza, and Gertjan van Noord. 2022. Evaluating pre-training objectives for low-resource translation into morphologically rich languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4933–4943, Marseille, France. European Language Resources Association.

Harshita Diddee, Kalika Bali, Monojit Choudhury, and Namrata Mukhija. 2022. The six conundrums of building and deploying language technologies for social good. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, COMPASS '22, page 12–19, New York, NY, USA. Association for Computing Machinery.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online. Association for Computational Linguistics.

Saket Dingliwal, Shuyang Gao, Sanchit Agarwal, Chien-Wei Lin, Tagyoung Chung, and Dilek Hakkani-Tur. 2021. Few shot dialogue state tracking using meta-learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1730–1739, Online. Association for Computational Linguistics.

Cheikh M. Bamba Dione, Alla Lo, Elhadji Mamadou Nguer, and Sileye Ba. 2022. Low-resource neural machine translation: Benchmarking state-of-the-art transformer for Wolof<->French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6654–6661, Marseille, France. European Language Resources Association.

Suma Reddy Duggenpudi, Subba Reddy Oota, Mounika Marreddy, and Radhika Mamidi. 2022. TeluguNER: Leveraging multi-domain named entity recognition with deep transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 262–272, Dublin, Ireland. Association for Computational Linguistics.

Abteen Ebrahimi, Arya D. McCarthy, Arturo Oncevay, John E. Ortega, Luis Chiruzzo, Gustavo Giménez-Lugo, Rolando Coto-Solano, and Katharina Kann. 2023. Meeting the needs of low-resource languages: The value of automatic alignments via pretrained models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3912–3926, Dubrovnik, Croatia. Association for Computational Linguistics.

Tobias Eder, Viktor Hangya, and Alexander Fraser. 2021. Anchor-based bilingual word embeddings for low-resource languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 227–232, Online. Association for Computational Linguistics.

Ramy Eskander, Francesca Callejas, Elizabeth Nichols, Judith Klavans, and Smaranda Muresan. 2020a. MorphAGram, evaluation and framework for unsupervised morphological segmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7112–7122, Marseille, France. European Language Resources Association.

Ramy Eskander, Smaranda Muresan, and Michael Collins. 2020b. Unsupervised cross-lingual part-of-speech tagging for truly low-resource scenarios. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP)*, pages 4820–4831, Online. Association for Computational Linguistics.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 587–593, Vancouver, Canada. Association for Computational Linguistics.

Hongliang Fei and Ping Li. 2020. Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5759–5771, Online. Association for Computational Linguistics.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anne Ferger. 2020. Processing language resources of under-resourced and endangered languages for the generation of augmentative alternative communication boards. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2644–2648, Marseille, France. European Language Resources Association.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2022. Dynamic gazetteer integration in multilingual models for cross-lingual and cross-domain named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2777–2790, Seattle, United States. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Fitsum Gaim, Wonsuk Yang, Hancheol Park, and Jong Park. 2023. Question-answering in a low-resourced language: Benchmark dataset and models for Tigrinya. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11857–11870, Toronto, Canada. Association for Computational Linguistics.

Andargachew Mekonnen Gezmu, Andreas Nürnberger, and Tesfaye Bayu Bati. 2022. Extended parallel corpus for Amharic-English machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6644–6653, Marseille, France. European Language Resources Association.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, S Ramaneswaran, S Sakshi, Utkarsh Tyagi, and Dinesh Manocha. 2023. DALE: Generative data augmentation for low-resource legal NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8511–8565, Singapore. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. ArXiv:2106.03193 [cs].

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

H Hammarström. 2009. . a survey of computational morphological resources for low-density languages. *NEALT*.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. 2022. Improving low-resource languages in pre-trained multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11993–12006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online. Association for Computational Linguistics.

William Held, Camille Harris, Michael Best, and Diyi Yang. 2023. A Material Lens on Coloniality in NLP. *arXiv*.

Marcelo Yuji Himoro and Antonio Pareja-Lora. 2022. Preliminary results on the evaluation of computational tools for the analysis of Quechua and Aymara. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5450–5459, Marseille, France. European Language Resources Association.

Yilun Hua, Zhaoyuan Deng, and Kathleen McKeown. 2023. Improving long dialogue summarization with semantic graph representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13851–13883, Toronto, Canada. Association for Computational Linguistics.

Yichong Huang, Xiaocheng Feng, Xinwei Geng, and Bing Qin. 2022. Unifying the convergences in multilingual neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6822–6835, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Ayyoob ImaniGooghari, Silvia Severini, Masoud Jalili Sabet, François Yvon, and Hinrich Schütze. 2022. Graph-based multilingual label propagation for low-resource part-of-speech tagging. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1577–1589, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Huiming Jin, Liwei Cai, Yihui Peng, Chen Xia, Arya McCarthy, and Katharina Kann. 2020. Unsupervised morphological paradigm completion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6696–6707, Online. Association for Computational Linguistics.

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.

Iman Jundi, Neele Falk, Eva Maria Vecchi, and Gabriella Lapesa. 2023. Node placement in argument maps: Modeling unidirectional relations in high & low-resource scenarios. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5854–5876, Toronto, Canada. Association for Computational Linguistics.

Huda Khayrallah, Brian Thompson, Matt Post, and Philipp Koehn. 2020. Simulated multiple reference training improves low-resource machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 82–89, Online. Association for Computational Linguistics.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.

Rebecca Knowles and Patrick Littell. 2022. Translation memories as baselines for low-resource machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6759–6767, Marseille, France. European Language Resources Association.

Wei-Jen Ko, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona Diab. 2021. Adapting high-resource NMT models to translate low-resource related languages without parallel data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 802–812, Online. Association for Computational Linguistics.

Boshko Koloski, Senja Pollak, Blaž Škrlj, and Matej Martinc. 2022. Out of thin air: Is zero-shot cross-lingual keyword detection better than unsupervised? In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 400–409, Marseille, France. European Language Resources Association.

Arjun Sai Krishnan and Seyoon Ragavan. 2021. Morphology-aware meta-embeddings for Tamil. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 94–111, Online. Association for Computational Linguistics.

Roland Kuhn. 2024. The Indigenous Languages Technology (ILT) project at the National Research Council of Canada, and its context - NRC Publications Archive - Canada.ca.

William Lane and Steven Bird. 2020. Interactive word completion for morphologically complex languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4600–4611, Barcelona, Spain (Online). International Committee on Computational Linguistics.

William Lane and Steven Bird. 2021a. Local word discovery for interactive transcription. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2058–2067, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

William Lane and Steven Bird. 2021b. Local Word Discovery for Interactive Transcription. *ACL Anthology*, pages 2058–2067.

Anna Langedijk, Verna Dankers, Phillip Lippe, Sander Bos, Bryan Cardenas Guevara, Helen Yannakoudakis, and Ekaterina Shutova. 2022. Meta-learning for fast cross-lingual adaptation in dependency parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8503–8520, Dublin, Ireland. Association for Computational Linguistics.

Leonard Collard Laura Stocker and Angela Rooney. 2016. Aboriginal world views and colonisation: implications for coastal sustainability†. *Local Environment*, 21(7):844–865.

Éric Le Ferrand, Steven Bird, and Laurent Besacier. 2022a. Learning From Failure: Data Capture in an Australian Aboriginal Community. *ACL Anthology*, pages 4988–4998.

Eric Le Ferrand, Steven Bird, and Laurent Besacier. 2022b. Learning from failure: Data capture in an Australian aboriginal community. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4988–4998, Dublin, Ireland. Association for Computational Linguistics.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8608–8621, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mingqi Li, Fei Ding, Dan Zhang, Long Cheng, Hongxin Hu, and Feng Luo. 2022a. Multi-level distillation of semantic knowledge for pre-training multilingual language model. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3097–3106, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shimin Li, Xiaotian Zhang, Yanjun Zheng, Linyang Li, and Xipeng Qiu. 2023a. Multijugate dual learning for low-resource task-oriented dialogue system. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11037–11053, Toronto, Canada. Association for Computational Linguistics.

Xiangyang Li, Xiang Long, Yu Xia, and Sujian Li. 2022b. Low resource style transfer via domain adaptive meta learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3014–3026, Seattle, United States. Association for Computational Linguistics.

Yu Li, Baolin Peng, Pengcheng He, Michel Galley, Zhou Yu, and Jianfeng Gao. 2023b. DIONYSUS: A pre-trained model for low-resource dialogue summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1368–1386, Toronto, Canada. Association for Computational Linguistics.

Shining Liang, Linjun Shou, Jian Pei, Ming Gong, Wanli Zuo, Xianglin Zuo, and Daxin Jiang. 2022. Label-aware multi-level contrastive learning for cross-lingual spoken language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9903–9918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Julian Linke, Philip N. Garner, Gernot Kubin, and Barbara Schuppler. 2022. Conversational speech recognition needs data? experiments with Austrian German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4684–4691, Marseille, France. European Language Resources Association.

Robert Litschko, Ivan Vulić, Željko Agić, and Goran Glavaš. 2020. Towards instance-level parser selection for cross-lingual transfer of dependency parsers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3886–3898, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Huadai Liu, Rongjie Huang, Xuan Lin, Wenqiang Xu, Maozong Zheng, Hong Chen, Jinzheng He, and Zhou Zhao. 2023a. ViT-TTS: Visual text-to-speech with scalable diffusion transformer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15957–15969, Singapore. Association for Computational Linguistics.

Ling Liu and Mans Hulden. 2020. Analogy models for neural word inflection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2861–2878, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zoey Liu, Justin Spence, and Emily Prud'hommeaux. 2023b. Investigating data partitioning strategies for crosslinguistic low-resource ASR evaluation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 123–131, Dubrovnik, Croatia. Association for Computational Linguistics.

Florian Lux and Thang Vu. 2022. Language-agnostic meta-learning for low-resource text-to-speech with articulatory features. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6858–6868, Dublin, Ireland. Association for Computational Linguistics.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2020. Tackling the low-resource challenge for canonical segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5237–5250, Online. Association for Computational Linguistics.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Christopher D. Manning, Kevin Jansz, and Nitin Indurkhya. 2001. Kirrkirr: Software for Browsing and Visual Exploration of a Structured Warlpiri Dictionary. *Lit. Linguist. Computing*, 16(2):135–151.

Kelly Marchisio, Ali Saad-Eldin, Kevin Duh, Carey Priebe, and Philipp Koehn. 2022. Bilingual lexicon induction for low-resource languages using graph matching via optimal transport. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2545–2561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Devansh Mehta, Sebastin Santy, Ramaravind Kommiya Mothilal, Brij Mohan Lal Srivastava, Alok Sharma, Anurag Shukla, Vishnu Prasad, Venkanna U, Amit Sharma, and Kalika Bali. 2020. Learnings from technological interventions in a low resource language: A case-study on Gondi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2832–2838, Marseille, France. European Language Resources Association.

Nelsi Melgarejo, Rodolfo Zevallos, Hector Gomez, and John E. Ortega. 2022. WordNet-QU: Development of a lexical database for Quechua varieties. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4429–4433, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nguyen Minh, Vu Hoang Tran, Vu Hoang, Huy Duc Ta, Trung Huu Bui, and Steven Quoc Hung Truong. 2022. ViHealthBERT: Pre-trained language models for Vietnamese in health text mining. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 328–337, Marseille, France. European Language Resources Association.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Sarah Moeller, Ling Liu, and Mans Hulden. 2021. To POS tag or not to POS tag: The impact of POS tags on morphological learning in low-resource settings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 966–978, Online. Association for Computational Linguistics.

Nikita Moghe, Evgeniia Razumovskaia, Liane Guillou, Ivan Vulić, Anna Korhonen, and Alexandra Birch. 2023. Multi3NLU++: A multilingual, multi-intent, multi-domain dataset for natural language understanding in task-oriented dialogue. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3732–3755, Toronto, Canada. Association for Computational Linguistics.

Syed Mostofa Monsur, Sakib Chowdhury, Md Shahrar Fatemi, and Shafayat Ahmed. 2022. SHONGLAP: A large Bengali open-domain dialogue corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5797–5804, Marseille, France. European Language Resources Association.

Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. An analysis of massively multilingual neural machine translation for low-resource languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3710–3718, Marseille, France. European Language Resources Association.

Jonathan Mukiibi, Andrew Katumba, Joyce Nakatumba-Nabende, Ali Hussein, and Joshua Meyer. 2022. The makerere radio speech corpus: A Luganda radio corpus for automatic speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1945–1954, Marseille, France. European Language Resources Association.

Saliha Muradoglu and Mans Hulden. 2022. Eeny, meeny, miny, moe. how to choose data for morphological inflection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7294–7303, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. *ACL Anthology*, pages 2144–2160.

Hellina Hailu Nigatu and Inioluwa Deborah Raji. 2024. "I Searched for a Religious Song in Amharic and Got Sexual Content Instead": Investigating Online Harm in Low-Resourced Languages on YouTube. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 141–160, New York, NY, USA. Association for Computing Machinery.

Rubungo Andre Niyongabo, Qu Hong, Julia Kreutzer, and Li Huang. 2020. KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5507–5521, Barcelona, Spain (Online). International Committee on Computational Linguistics.

NLLB. 2024. Scaling neural machine translation to 200 languages.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. KinyaBERT: a morphology-aware Kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363, Dublin, Ireland. Association for Computational Linguistics.

Bruce Oliver, Clarissa Forbes, Changbing Yang, Farhan Samir, Edith Coates, Garrett Nicolai, and Miikka Silfverberg. 2022. An inflectional database for gitksan. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6597–6606, Marseille, France. European Language Resources Association.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. BAD-X: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1791–1799, Seattle, United States. Association for Computational Linguistics.

Md Rizwan Parvez and Kai-Wei Chang. 2021. Evaluating the values of sources in transfer learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5084–5116, Online. Association for Computational Linguistics.

Dhrubajyoti Pathak, Sukumar Nandi, and Priyankoo Sarmah. 2022. AsNER - annotated dataset and baseline for Assamese named entity recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6571–6577, Marseille, France. European Language Resources Association.

Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233, Dublin, Ireland. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aidan Pine, Dan Wells, Nathan Brinklow, Patrick Littell, and Korin Richmond. 2022a. Requirements and motivations of low-resource speech synthesis for language revitalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7346–7359, Dublin, Ireland. Association for Computational Linguistics.

Aidan Pine, Dan Wells, Nathan Brinklow, Patrick Littell, and Korin Richmond. 2022b. Requirements and Motivations of Low-Resource Speech Synthesis for Language Revitalization. *ACL Anthology*, pages 7346–7359.

Edoardo M. Ponti, Ivan Vulić, Ryan Cotterell, Marinela Parovic, Roi Reichart, and Anna Korhonen. 2021. Parameter space factorization for zero-shot learning across tasks and languages. *Transactions of the Association for Computational Linguistics*, 9:410–428.

Namoos Hayat Qasmi, Haris Bin Zia, Awais Athar, and Agha Ali Raza. 2020. SimplifyUR: Unsupervised lexical text simplification for Urdu. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3484–3489, Marseille, France. European Language Resources Association.

Kunxun Qi, Hai Wan, Jianfeng Du, and Haolan Chen. 2022. Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1910–1923, Dublin, Ireland. Association for Computational Linguistics.

Annie Rajan, Ambuja Salgaonkar, and Ramprasad Joshi. 2020. A survey of konkani nlp resources. *Computer Science Review*, 38:100299.

Arun Ramachandran and Gerard de Melo. 2020. Cross-lingual emotion lexicon induction using representation alignment in low-resource settings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5879–5890, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sahana Ramnath, Melvin Johnson, Abhirut Gupta, and Aravindan Raghuveer. 2021. HintedBT: Augmenting Back-Translation with quality and transliteration hints. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1717–1733, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 African languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1306–1320, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shruti Rijhwani, Shuyan Zhou, Graham Neubig, and Jaime Carbonell. 2020. Soft gazetteers for low-resource named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8118–8123, Online. Association for Computational Linguistics.

Renato Rocha Souza, Amelie Dorn, Barbara Piringer, and Eveline Wandl-Vogt. 2020. Identification of indigenous knowledge concepts through semantic networks, spelling tools and word embeddings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 943–947, Marseille, France. European Language Resources Association.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Efsun Sarioglu Kayi, Linyong Nan, Bohan Qu, Mona Diab, and Kathleen McKeown. 2020. Detecting urgency status of crisis tweets: A transfer learning approach for low resource languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4693–4703, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kumar Saunack, Kumar Saurav, and Pushpak Bhattacharyya. 2021. How low is too low? a monolingual take on lemmatisation in Indian languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4088–4094, Online. Association for Computational Linguistics.

Kumar Saurav, Kumar Saunack, and Pushpak Bhattacharyya. 2020. Analysing cross-lingual transfer in lemmatisation for Indian languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6070–6076, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Michael Schlichtkrull and Anders Søgaard. 2017. Cross-lingual dependency parsing with late decoding for truly low-resource languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 220–229, Valencia, Spain. Association for Computational Linguistics.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. *ACL Anthology*, pages 724–731.

Harshita Sharma, Pruthwik Mishra, and Dipti Sharma. 2022. HAWP: a dataset for Hindi arithmetic word problem solving. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3479–3490, Marseille, France. European Language Resources Association.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, Online. Association for Computational Linguistics.

Gerardo Sierra Martínez, Cynthia Montaño, Gemma Bel-Enguix, Diego Córdova, and Margarita Mota Montoya. 2020. CPLM, a parallel corpus for Mexican languages: Development and interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2947–2952, Marseille, France. European Language Resources Association.

Claytone Sikasote and Antonios Anastasopoulos. 2022. BembaSpeech: A speech recognition corpus for the Bemba language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7277–7283, Marseille, France. European Language Resources Association.

Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. 2022. Assessing digital language support on a global scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Heather Simpson, Christopher Cieri, Kazuaki Maeda, Kathryn Baker, and Boyan Onyshkevych. 2008. Human Language Technology Resources for Less Commonly Taught Languages: Lessons Learned Toward Creation of Basic Language Resources. *ACL Anthology*.

Alexey Sorokin. 2020. Getting more data for low-resource morphological inflection: Language models and data augmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3978–3983, Marseille, France. European Language Resources Association.

Daniel Stiles. 1982. A history of the hunting peoples of the northern east africa coast: Ecological and socio-economic considerations. *Paideuma*, 28:165–174.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676, Dublin, Ireland. Association for Computational Linguistics.

Haoran Sun and Deyi Xiong. 2022. Language branch gated multilingual neural machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5046–5053, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. 2022a. Multimodal Dialogue Response Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2854–2866, Dublin, Ireland. Association for Computational Linguistics.

Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. 2022b. Multimodal dialogue response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2854–2866, Dublin, Ireland. Association for Computational Linguistics.

Mauro Tosco. 1992. Dahalo: an Endangered Language. *Language Death: Factual and Theoretical Explorations ("Contributions to the Sociology of Language 64")*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online. Association for Computational Linguistics.

Jessica Wakeman. 2021. Cherokee fight to save language from extinction.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Fei Huang, and Kewei Tu. 2020. Structure-level knowledge distillation for multilingual sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3317–3330, Online. Association for Computational Linguistics.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

Mengzhou Xia, Guoqing Zheng, Subhabrata Mukherjee, Milad Shokouhi, Graham Neubig, and Ahmed Hassan Awadallah. 2021. MetaXL: Meta representation transformation for low-resource cross-lingual learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 499–511, Online. Association for Computational Linguistics.

Yi Xu, Shuqian Sheng, Jiexing Qi, Luoyi Fu, Zhouhan Lin, Xinbing Wang, and Chenghu Zhou. 2023. Unsupervised graph-text mutual conversion with a unified pretrained language model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5130–5144, Toronto, Canada. Association for Computational Linguistics.

Isil Yakut Kilic and Shimei Pan. 2022. Incorporating LIWC in neural networks to improve human trait and behavior analysis in low resource scenarios. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4532–4539, Marseille, France. European Language Resources Association.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ayele, and Chris Biemann. 2020. Exploring Amharic sentiment analysis from social media texts:

Building annotation tools and classification models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1048–1060, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Michelle Yuan, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan Boyd-Graber. 2020. Interactive refinement of cross-lingual word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5984–5996, Online. Association for Computational Linguistics.

Jiali Zeng, Yufan Jiang, Yongjing Yin, Yi Jing, Fandong Meng, Binghuai Lin, Yunbo Cao, and Jie Zhou. 2023. Soft language clustering for multilingual model pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7021–7035, Toronto, Canada. Association for Computational Linguistics.

Rodolfo Zevallos and Nuria Bel. 2023a. Hints on the data for language modeling of synthetic languages with transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12508–12522, Toronto, Canada. Association for Computational Linguistics.

Rodolfo Zevallos and Núria Bel. 2023b. Hints on the data for language modeling of synthetic languages with transformers. *ACL Anthology*, pages 12508–12522.

Rui Zhang, Caitlin Westerfield, Sungrok Shim, Garrett Bingham, Alexander Fabbri, William Hu, Neha Verma, and Dragomir Radev. 2019. Improving low-resource cross-lingual document retrieval by reranking with deep bilingual representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3173–3179, Florence, Italy. Association for Computational Linguistics.

Guolin Zheng, Yubei Xiao, Ke Gong, Pan Zhou, Xiaodan Liang, and Liang Lin. 2021. Wav-BERT: Cooperative acoustic and linguistic representation learning for low-resource speech recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2765–2777, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shuyan Zhou, Shruti Rijhwani, John Wieting, Jaime Carbonell, and Graham Neubig. 2020. Improving candidate generation for low-resource cross-lingual entity linking. *Transactions of the Association for Computational Linguistics*, 8:109–124.

Yicheng Zou, Bolin Zhu, Xingwu Hu, Tao Gui, and Qi Zhang. 2021. Low-resource dialogue summarization with domain-agnostic multi-source pretraining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 80–91, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Labels for Classifying Languages

In this section, we provide the descriptions for labels used for language vitality and digital access used in Figure 4.

A.1 Vitality

In this work, we refer to the scale from Ethnologue[14] which is derived from the Expanded Graded Intergenerational Disruption Scale (EGIDS) (Anderbeck, 2015).

Institutional — The language has been developed to the point that it is used and sustained by institutions beyond the home and community.

Stable — The language is not being sustained by formal institutions, but it is still the norm in the home and community that all children learn and use the language.

Endangered — It is no longer the norm that children learn and use this language.

Extinct - The language is no longer used, and no one retains a sense of ethnic identity associated with the language.

A.2 Digital Access

This taxonomy is from Simons et al. (2022) and is also used by Ethnologue.

Still — this language shows no signs of digital support

Emerging — the language has some content in digital form and/or encoding tools

Ascending — the language has some spell checking or localized tools or machine translation as well

Vital — the language is supported by multiple tools in all of the above categories and as well as some speech processing

Thriving — the language has all of the above plus virtual assistants

B Criteria used in Studying Languages

Figure 5 shows the distributions of the various criteria used for categorizing a language as 'low-resource' in the studied languages. Figure 6 depicts

[14] https://www.ethnologue.com/

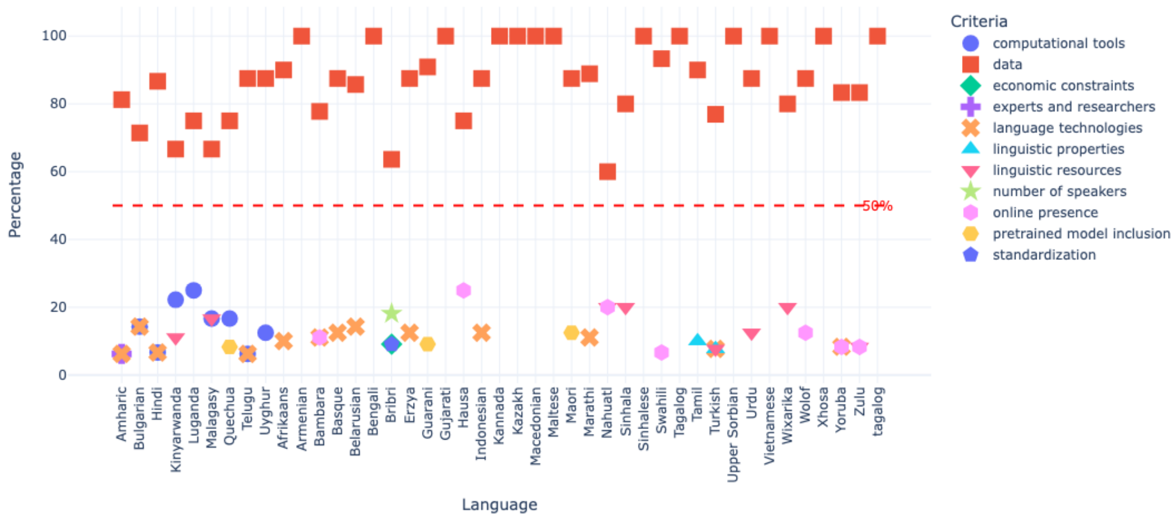

Figure 5: Distribution of criteria stated by papers in our study to categorize languages as low-resource.

different perspectives used to refer to the lack of a dataset for a language.

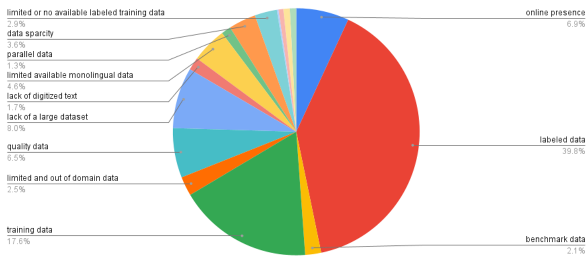

Figure 6: Criteria used in the papers to show lack of data.

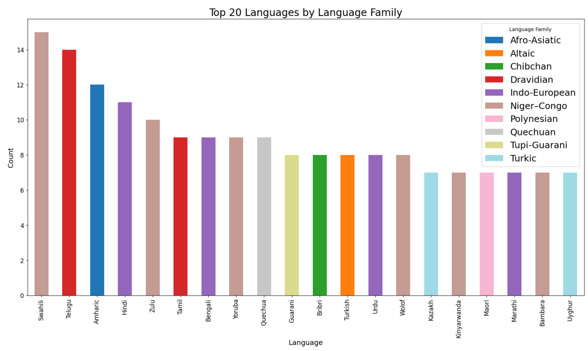

Figure 7: Number of papers per language for the top-20 most studied languages.

C Most frequently studied languages

Figure 7 shows the top 20 most frequently studied languages in our sample. We see that Swahili and Telugu take the lead with 14 papers working on them. Geographically, we observe that Indian languages ($n = 7$) are the most represented in our sample, with an equal number of languages ($n = 7$) from the entire continent of Africa.

D Categories used to define low-resource

Here, we grouped papers according to the criteria used in the paper to categorize a language as a low-resource language.

Socio-political [(Maillard et al., 2023; Coto-Solano, 2022; Pathak et al., 2022)]

Resources

Native Speakers [(Pine et al., 2022a; Oliver et al., 2022; Coto-Solano, 2022; Feldman and Coto-Solano, 2020; Leong et al., 2022)]

Online Presence [(Bamutura et al., 2020; Sierra Martínez et al., 2020; Adelani et al., 2022; Nzeyimana and Niyongabo Rubungo, 2022; Feldman and Coto-Solano, 2020; Bustamante et al., 2020; Patil et al., 2022; Adelani et al., 2022)]

Language experts [(Brixey et al., 2020; Yimam et al., 2020)]

Artifacts

Linguistic Knowledge [(Qasmi et al., 2020; Coto-Solano, 2022)]

Data [Ferger (2020); Zevallos and Bel (2023a); Pine et al. (2022a); Fei and Li (2020); Eskander et al. (2020a); Xia et al. (2021); Goyal et al. (2022); Sorokin (2020); Pfeiffer et al. (2020); Sierra Martínez et al. (2020); Ahuja et al. (2022); Mehta et al. (2020); Le Ferrand et al. (2022b); Mukiibi et al. (2022); Chaudhary et al. (2021); Üstün et al. (2020); Eskander et al. (2020b); Liang et al. (2022); Pfeiffer et al. (2021); ImaniGooghari et al. (2022); Dione et al. (2022); Chukwuneke et al. (2022); Schmidt et al. (2022); Hasan et al. (2020); Muradoglu and Hulden (2022); Biswas et al. (2020); Marchisio et al. (2022); Maillard et al. (2023); Litschko et al. (2020); Coto-Solano (2022); Gaim et al. (2023); Adebara et al. (2022); Krishnan and Ragavan (2021); Alabi et al. (2020); Yimam et al. (2020); Li et al. (2022a); Saunack et al. (2021); Niyongabo et al. (2020); Ramnath et al. (2021); Ponti et al. (2021); Adouane et al. (2020); Reid et al. (2021); Parović et al. (2022); Minixhofer et al. (2022); Zeng et al. (2023); Pathak et al. (2022); Botha et al. (2020); Chakrabarty et al. (2022); Debnath et al. (2021); Sarioglu Kayi et al. (2020); Alabi et al. (2022); Ko et al. (2021); Liu and Hulden (2020); Wang et al. (2020); Zhou et al. (2020); Sharma et al. (2022); Bari et al. (2021); ImaniGooghari et al. (2023); Yuan et al. (2020); Gezmu et al. (2022); Qi et al. (2022); Knowles and Littell (2022); Khayrallah et al. (2020); Mager et al. (2020); Monsur et al. (2022); Ramachandran and de Melo (2020); Sun and Xiong (2022); Hangya et al. (2022); Saurav et al. (2020); Ouyang et al. (2021); Parvez and Chang (2021); Moeller et al. (2021); Fomicheva et al. (2022); Mueller et al. (2020); Siddhant et al. (2020); Bartelds et al. (2023); Daniel et al. (2019); Chen et al. (2022); Fetahu et al. (2022); Li et al. (2022b,b); Bartelds and Wieling (2022); Minixhofer et al. (2022); Minh et al. (2022); Koloski et al. (2022); Coto-Solano et al. (2022); Yakut Kilic and Pan (2022); Linke et al. (2022); Langedijk et al. (2022); Muradoglu and Hulden (2022); Huang et al. (2022); Jundi et al. (2023); Xu et al. (2023); Li et al. (2023a); Su et al. (2022); Hua et al. (2023); Li et al. (2023b); Sun et al. (2022b); Moghe et al. (2023); Bhat et al. (2023); de Vries et al. (2022); Eder et al. (2021); Zhang et al. (2019); Fang and Cohn (2017); Xia et al. (2019); Liu et al. (2023b); Schlichtkrull and Søgaard (2017); Dingliwal et al. (2021); Ebrahimi et al. (2023); Röttger et al. (2022); Ghosh et al. (2023); Ding et al. (2020); Zou et al. (2021); Lux and Vu (2022); Zheng et al. (2021); Liu et al. (2023a)]

Technology [(Bamutura et al., 2020; Byamugisha, 2022; Melgarejo et al., 2022; Yimam et al., 2020; Himoro and Pareja-Lora, 2022; Li et al., 2022a; Niyongabo et al., 2020; Duggenpudi et al., 2022; Avram et al., 2022; Lane and Bird, 2021b; Eskander et al., 2020a; Rocha Souza et al., 2020; Lane and Bird, 2020; de Lhoneux et al., 2022; ImaniGooghari et al., 2022; Brixey et al., 2020; Rijhwani et al., 2020; Sikasote and Anastasopoulos, 2022; Adouane et al., 2020; Botha et al., 2020; Moeller et al., 2021; Jin et al., 2020; Dhar et al., 2022; Pfeiffer et al., 2020; Leong et al., 2022)]