Building Better: Avoiding Pitfalls in Developing Language Resources when Data is Scarce

Anonymous ACL submission

Abstract

Language is a symbolic capital that affects people's lives in many ways (Bourdieu, 1977, 1991). It is a powerful tool that accounts for identities, cultures, traditions, and societies in general. Hence, data in a given language should be viewed as more than a collection of tokens. Good data collection and labeling practices are key to building more human-centered and socially aware technologies. While there has been a rising interest in mid- to low-resource languages within the NLP community, work in this space has to overcome unique challenges such as data scarcity and access to suitable annotators. In this paper, we collect feedback from those directly involved in and impacted by NLP artefacts for mid- to low-resource languages. We conduct a quantitative and qualitative analysis of the responses and highlight the main issues related to (1) data quality such as linguistic and cultural data suitability; and (2) the ethics of common annotation practices such as the misuse of online community services. Based on these findings, we make several recommendations for the creation of high-quality language artefacts that reflect the cultural milieu of its speakers, while simultaneously respecting the dignity and labor of data workers.

1 Introduction

There has been an increasing interest in improving the current scope of NLP research with more human-centered design choices (Kotnis et al., 2022) and the inclusion of social awareness (Yang et al., 2024) and underrepresented world populations (Mihalcea et al., 2024). As language technologies depend on the quality of the data (Hirschberg and Manning, 2015) and their alignment with the needs of the speakers, researchers, and other users, the perspectives of these different stakeholders are key to high-quality tools and resources. That is, data selection, annotation, and design choices are traditionally made by the researchers who develop

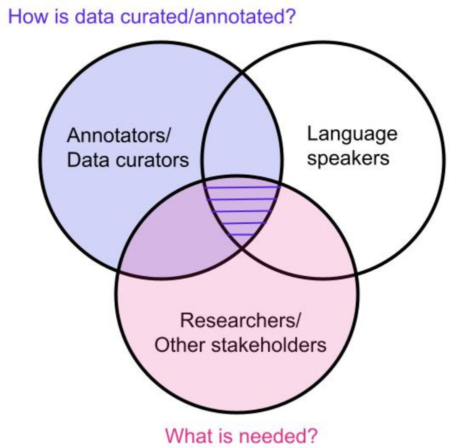

Figure 1: The main themes and targets of our survey. It is designed for NLP researchers and practitioners who have worked on non-high-resource languages (data curation, annotation, and/or model construction). Some of the questions focus on the perspectives of the subset highlighted in the figure, i.e., speakers who focus on their own languages.

the different artefacts. However, the involvement of those whose native languages are in question is paramount to better design practices (Bird and Yibarbuk, 2024) as language is part of their culture and identity (Bourdieu, 1991). That said, when dealing with mid- to low-resource languages in NLP, researchers often make use of the datasets available without necessarily looking into their adequacy, mainly due to resource scarcity. Although progress in NLP for English and other high-resource languages has led to improving standards for corpora quality control and research practices (Gebru et al., 2021; Bender and Friedman, 2018; Mohammad, 2022), one cannot claim the same about the data sources and prevailing practices for mid- to low-resource languages given the current research scope in the field (Joshi et al., 2020). Therefore, the NLP artefacts developed for low-resource languages and underrepresented cultures often suffer from a lack of social considerations and overgeneralisations due to the over-reliance on data and tools that fail to incorporate the predominant linguistic and cultural features of a given language (Bender and Friedman, 2018), which may hinder critical progress. This can further lead to inequality (Blasi et al., 2022; Held et al., 2023), sub-optimal experiences with language technologies, and could reinforce a legacy of language hierarchy (Kahane, 1986).

In this position paper, we shed light on the current limitations of NLP research for mid- to low-resource languages in terms of appropriate data collection, ethical annotation practices, and overall data quality. We reached out to the NLP community involved in NLP projects on under-served languages and conducted a survey to report on the common incentives, limitations, applied norms, and practices (see Figure 1). We outline the survey and present its results. Finally, based on the survey responses, we provide a set of recommendations that focus on (1) fairness and centering of the speakers of the language, (2) choosing suitable data sources, (3) setting fair and realistic expectations when recruiting annotators, and (4) avoiding cultural misrepresentation.

2 Related Work

Work on ethical practices in AI, ML, and NLP research covers a variety of topics, such as artefact documentation (Bender, 2011; Bender and Friedman, 2018; Gebru et al., 2021; Rogers et al., 2021; Mohammad, 2022) and recommendations for best practices (Hollenstein et al., 2020; Mohammad, 2023). Those that focus on low-resource languages are centered on the general state of NLP research in the area (Held et al., 2023; Joshi et al., 2020; Blasi et al., 2022; Doğruöz and Sitaram, 2022), limitations in specific tasks such as machine translation (Mager et al., 2023), LLM research (Mihalcea et al., 2024), or on the essential question of including people whose languages are in question (Mager et al., 2023; Bird, 2020, 2022; Bird and Yibarbuk, 2024; Lent et al., 2022). Such work sheds light on the peculiarities of low-resource languages with the majority being vernacular languages rather than institutionalised or written (Bird and Yibarbuk, 2024; Bird, 2024). They further advocate for language communities to take over their languages (Schwartz, 2022; Markl et al., 2024; Mihalcea et al., 2024). For instance, Bird and Yibarbuk (2024) focus on how experts (e.g., linguists, computer scientists) interact with the language communities using participatory design approaches (Winschiers-Theophilus et al., 2010), and Cooper et al. (2024) provide recommendations on how to engage with indigenous communities without merely focusing on accuracy. Doğruöz and Sitaram (2022) further point out the importance of not treating language technologies for low-resource languages as scaled-down versions of high-resource ones, and Adebara and Abdul-Mageed (2022) make similar claims with a focus on features that are specific to African languages. In addition to the language speakers, other work focuses on users such as Blaschke et al. (2024) who highlight the needs of dialect speakers and the importance of involving end users in designing language technologies. Moreover, Yang et al. (2024) define social awareness and advocate for refraining from treating language in NLP as a computational problem only. In this paper, we strengthen the above discussion by shifting the focus to the practical challenges faced by NLP researchers and practitioners working on mid- to low-resource languages by borrowing practices from social science (Cetina, 1999) to study the methodological practices and issues in the field. To the best of our knowledge, there is limited work investigating NLP research for low-resource languages while trying to connect to online communities, except for three case studies discussed by Birhane et al. (2022), and work by (Lent et al., 2022), who analyse 38 responses collected on Facebook and Twitter. By analysing the respondents' feedback, we aim to present practical recommendations that emphasise transparency and ethically grounded practices for building more human-centered NLP artefacts for mid- to low-resource languages.

3 Survey

Our main goal is to investigate the current issues and problematic practices in NLP research for mid- to low-resource languages and provide potential solutions. Therefore, we reached out to the NLP community from June to October 2024 on X, LinkedIn, Google groups and Slack channels of NLP communities, and by direct emails. We targeted researchers working on mid- and low-resource languages, language variants, dialects, and vernaculars, and surveyed how research is conducted. Participants report on common practices, incentives, and issues that stand out. Then, we present a quantitative and qualitative analysis of the responses.

3.1 Respondents

Respondents are NLP researchers and practitioners involved in the data collection, annotation, model construction, or other research questions related to mid- to low-resource languages. Some may have also conducted research for high-resource languages. Note that the respondents may or may not speak the language(s).

3.2 Survey Structure

We ask the respondents about (1) their previous experiences in the area, (2) current problems and limitations relevant to their language(s) of interest, (3) the motivation behind their involvement in various projects, and (4) how they were credited for tasks that are often specific to low-resource languages, e.g., compensation for annotations done via online community forums. Note that we left it to the respondents to decide on what represents a mid- to low-resource language.

3.2.1 General Questions

Respondents could optionally fill in their names and contact information for a potential follow-up. Then, they were asked about:

- the language(s) they work on,
- the project(s) they were involved in,
- whether they are/were part of any online community,
- whether the project(s) they worked on are from industry, academia, or both,
- the kind of NLP tools that are or would be relevant and useful in their language(s) on interest,
- the reason(s) why they work on this/these language(s).

3.2.2 Reporting on Incentives and Potential Limitations

We investigate the common reasons why researchers work on low-resource languages. Therefore, we ask the participants to report on:

- the incentive(s) for working on their language(s) of interest,
- the incentive(s) for working on specific projects.

As we are aware of potential drawbacks in NLP for mid- to low-resource languages (Blasi et al., 2022), we examined whether the respondents work in the area due to any limitations observed in available NLP tools in their language(s) of interest. Note that these questions were optional as researchers may work on mid- to low-resource languages for various other reasons. We asked the participants to report on:

- any observed limitations and optionally list some tools or resources in their language(s) of interest as examples,
- potential language-specific challenges in their language(s) of interest.

3.2.3 Reporting on Credit Attribution

We asked the respondents about how often they were properly credited for their work. Further, as reaching out to online communities is common to projects that include mid- to low-resource languages, we asked whether the participants were involved in past projects through online community platforms (for data collection, annotation, model construction, etc.). This is because involving communities in NLP and ML projects is relatively new to the field and can therefore be abused as there are no clear standards regarding data workers in such contexts. Therefore, our questions were the following:

- How often did the respondents receive credit for their contributions? E.g., whether they received proper financial compensation for annotating a dataset.
- How often were they offered authorship when making substantial contributions to the data collection and/or data annotation?
- What were their incentives for projects in which they did not receive financial compensation or authorship?
- How long did the process take especially when they were not properly compensated?

4 Findings

We received 81 responses from researchers working on a wide range of mid- to low-resource languages and language families. Even though including contact information was optional, more than 90% of the respondents chose not to reply anonymously, and 80% asked for updates on the project. Table 1 shows the distribution of responses to questions on project affiliations, the tasks in which the respondents were involved, and their motivations for working on mid- to low-resource languages. Note that percentages do not sum up to one as respondents could report on more than one project.

Projects in		Task		Motivation	
Industry	12%	Data creation	47%	Scientific interest	81%
Academia	57%	Data annotation	33%	Building language technologies	72%
Both	31%	Data collection	33%	Limitations in language(s) of interest	60%
		Model construction	9%	LLM research	59%

Table 1: Reported project affiliations, tasks in which the annotators were involved, and the different motivations or incentives. Note that percentages do not sum up to one as respondents could report on more than one project.

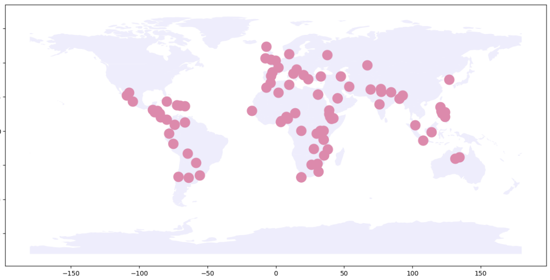

Figure 2: The main locations where the languages that our survey respondents work on are spoken.

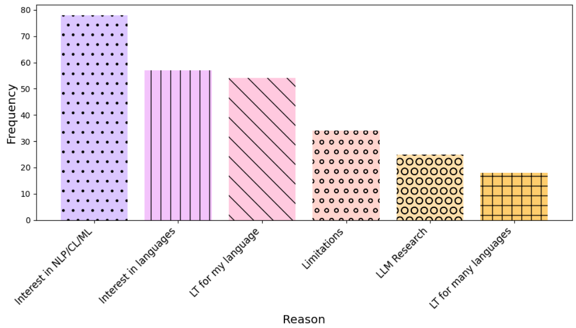

Figure 3: Frequency of each incentive that was found in our survey responses. Note that the percentages do not sum up to 100 as the respondents could choose more than one option.

That is. participants could be involved in several tasks and projects. As shown in Table 1, most participants were involved in dataset curation mainly motivated by scientific interest or curiosity, and for building language technologies because of observed limitations in resources dedicated to their language(s) of interest.

4.1 General Information

4.1.1 Projects

The respondents could report on one or many projects they have been involved in. As shown in Table 1, Most respondents have worked on academic projects, with a third on collaborations between industry and academia or both types of projects.[1]

4.1.2 Languages

Among the 81 responses, the respondents worked on >70 low-resource languages they specifically named (see Appendix). Figure 2 illustrates the main locations where these languages are spoken. The languages include variants, dialects, and vernaculars (e.g., country-specific Arabic dialects), truly low-resource languages (e.g., Welsh, Yoreme Nokki, Setswana), and mid- to low-resource ones (e.g., Amharic, Indonesian). In addition, about 12% of the respondents reported working on language families and language branches such as South Asian languages, all Gaeilige dialects, or Arabic/English variations. A high percentage of the respondents work on high-resource languages as well, such as English, French, Spanish, and Modern Standard Arabic.

[1]Note that although >50% of the respondents named the projects they participated in and did not mind sharing this information publicly, we do not disclose it to protect the anonymity of our respondents.

4.2 Incentives and Potential Limitations

When asking the respondents about why they work on NLP for mid- to low-resource languages, we provide them with a checklist from which they could choose more than one option or add their own entry. We report on the frequent motivations and practices that are only adopted in non-high-resource settings due to, e.g., data scarcity. We identify problematic instances and analyse the possible reasons behind some. When further examining the common motivations, we report more detailed numbers in Figure 3. Among those who were motivated by scientific curiosity or interest in Table 1 there were those whose interest was in NLP/CL/ML research (68%) and those whose interest was in languages (68%). Note that the two are not mutually exclusive.

Moreover, for the respondents whose motiva-

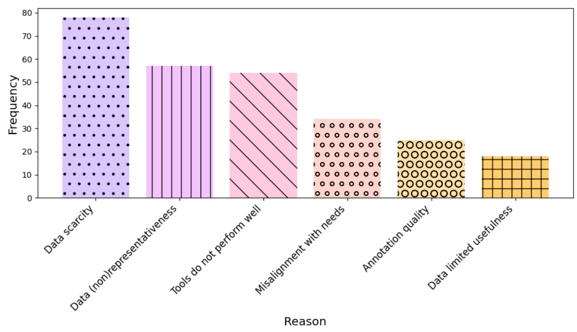

Figure 4: Frequency of each reported limitation when the respondents reported working on NLP for low-resource languages due to marked shortcomings.

tion was building language technologies, most of them were more interested in building technologies for their own language(s) (60%) as opposed to building technologies for as many languages as possible (38%). This is particularly interesting as it constitutes evidence of the power of language as a symbolic capital (Bourdieu, 1991), which can sometimes manifest in the feeling of "a duty" that one has towards their language. Other frequent motivations include marked limitations in language resources and tools in the language(s) of interest (60%) and the willingness to contribute to research on LLMs (59%).

4.2.1 Reported Limitations

More than 60% of the respondents reported working on low-resource languages due to marked limitations in currently available resources for their language(s) of interest. To shed light on these limitations, we showed the respondents a predefined list of possible shortcomings as well as a text box where they could add any observed limitations. As shown in Figure 4: the predominant limitation is data scarcity (78%). This is followed by the lack of representativeness of the data (58%), the underperformance of the available tools (54%), their misalignment with the users' needs (54%), the low quality of the annotations (25%), and the lack of the usefulness of the data (18%).

4.2.2 Qualitative Analysis of the Limitations

We provided the respondents with free text sections where they could report examples of tools or resources that suffer from the limitations that they mentioned to justify their choices. When manually processing the answers, we noticed the following themes:

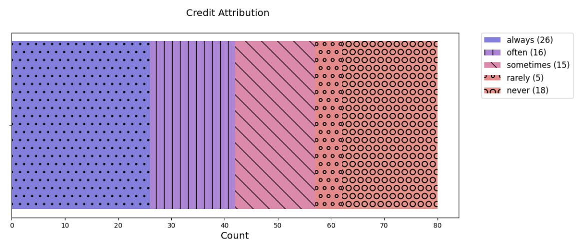

Figure 5: Respondents on getting credit for projects they were involved in.

1. **Limitations related to the currently available resources:** such as their unavailability, small size, limited representativeness, and quality.
2. **Limitations related to the practices adopted when building new resources:** such as:
    - the reliance on machine translation tools and LLMs to build resources for under-resourced languages;
    - the lack of awareness of culture-specific and linguistic challenges of the languages in question;
    - the challenges with annotator recruitment due to the lack of availability of native or near-native speakers on commonly used annotation platforms (e.g., AMT and Prolific),
    - the potential misuse of online community services.
3. **Fundamental problems related to NLP research on mid- to low-resource languages:** such as the lack of funding often due to the "low prestige" language dilemma—the false notion that some languages or language varieties are more important than others.

We discuss all three of these themes below.

Currently Available Resources As many languages are not institutional but rather vernacular (Bird and Yibarbuk, 2024), data collection presents considerable challenges when solely relying on textual data, e.g., Bantu languages.

Further, the focus on English and the reliance on translated data harms the quality of the generated datasets as they do not capture the subtle peculiarities of a given language. Another issue is what is commonly called "the curse of multilinguality" as the commonly used multilingual tools do not perform as well as the monolingual ones. It is important to note that what is translated and whether it was further verified by a native speaker makes a difference. For instance, translating Wikipedia texts can be easier than translating conversational, informal, or religious texts (Hutchinson, 2024).

Limitations with respect to Building New Resources Lack of representativeness and naturalness as well as "attention to details" were commonly reported in the responses. The respondents reported a lack of awareness of language variants and cultural aspects when building a language-specific artefact; the reliance on the standardised version of a given language due to power dynamics (more power in the hands of well-funded institutions and established researchers); the presence of offensive utterances in the data due to a lack of data filtering; and potentially wrong assumptions about a language or a culture. Further, the time-specific context and usage of some languages, such as ancestral ones (e.g., Coptic), have considerably changed and one has to take these facts into account. In addition, datasets may be collected from inadequate sources or could be aligned with Western values, standards, or expectations. This can be due to power differentials or a lack of deeper examination carried along with locals and native speakers. Finally, researchers rely on personal connections as it is hard to impossible to find native speakers of mid- to low-resource languages on commonly used annotation platforms such as Amazon Mechanical Turk, Polific, and others. Added to this reason, the lack of funding leads researchers to turn to online community work. This practice has been at the center of major NLP contributions in recent years (Birhane et al., 2022). However, despite its benefits for people with common research interests and incentives, the absence of well-established standards puts community members at risk as their efforts may not be properly recognised.

Fundamental Problems Further, many respondents reported that conducting research in mid- to low-resource languages often entailed high costs of data curation, potential reach out to local communities, the need for resources, and the cost of the datasets that are not freely available.

4.3 Credit Attribution

We asked the respondents to share whether they were properly credited for their work by, e.g., getting financial compensation for a long annotation task, getting involved in the writing of a research paper for a resource that they built, etc. As shown in Figure 5, most respondents (>67%) report this

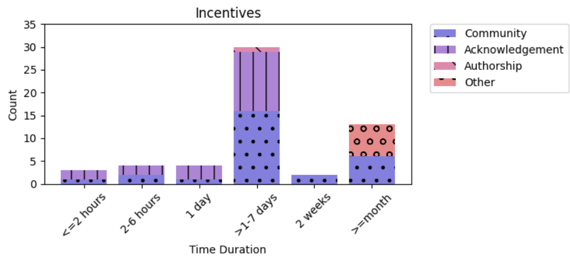

Figure 6: Respondents on incentives when no proper credit (e.g., financial compensation for data annotation) was offered. We show the counts of various incentives and the time it took the participants to complete their work for a given project (from <=2 hours to more than a month).

not being the case at least once. Figure 6 shows the distributions of responses pertaining to how the respondents were incentivised to perform an annotation task for which they were eventually not given due credit.

Problematic Incentivisation For the respondents who reported that they did not receive proper credit for at least one project they were involved in, we report the initial incentives for joining these projects and the time it took the participants to complete the work. As shown in Figure 6, they were either:

1. a member of a community (see paragraph below), or
2. acknowledged on the website or the research paper, or
3. somehow manipulated into thinking that there was a professional benefit in joining without proper compensation.

The Issue with the Over-reliance on Online Communities When using standard crowdsourcing platforms such as AMT or Prolific, one can operationalise the annotation for a given task. Despite their shortcomings (Fort et al., 2011; Irani, 2015), one can attempt to protect workers by using tests and training when annotating hard tasks. However, for mid- to low-resource languages, platforms such as AMT and Prolific often do not have enough speakers registered on the platform. Therefore, researchers opt for personal connections or community efforts instead. There are various advantages to personal outreach and community efforts, such as the fact that people feel more included and trust can be built more easily. On the other hand, there is a high risk of exploitation and emotional manipulation in such a case, junior researchers can be told that joining an online community that helps build resources for a language is prestigious and worth adding to their CVs. We note that some respondents shared their frustration in the responses. As shown in Figure 6, 40% of the respondents, who spent 1 day to more than a month annotating data report negative experiences. That is, their work was not properly compensated, acknowledged, or recognised. This calls for a need to set guidelines and standards when using community services.

5 Recommendations

While there has been a considerable amount of work on the ethics of best practices for building NLP and ML artefacts (Bender and Friedman, 2018; Leech et al., 2024; Mohammad, 2022, 2023), our findings substantiate the fact that research on mid- to low-resource languages presents additional challenges linked to the reliance on unconventional practices. While we do not expect the datasets to be perfect, one can address the most pressing issues and report the remaining ones in the limitations section of a resource paper.

5.1 Center the People

Our findings show that there are various issues that ought to be addressed early as research in the area lacks established standards and is subject to power differentials. Many mid- to low-resource languages are from what is called "the Global South" with a large number of them being spoken rather than written.

Speakers Language is an important part of a population's identity and technologies dealing with it have a direct impact on people's lives. Past NLP work highlights how to engage with speakers and communities whose languages are in question (Bird and Yibarbuk, 2024; Bird, 2020, 2024; Cooper et al., 2024). We further reinforce this argument with our findings.

When a researcher reaches out to a group with little background knowledge of their culture or language, one needs to approach these problems from the perspective of the community in question (Bird, 2022). Hence, the question of **who is exactly served** needs to be addressed early on to avoid any misconception of perceived needs for language technologies.

Researchers vs. Data Workers In addition to the large percentage of our survey respondents who reported not being properly credited for their labor, there were cases of emotional manipulation (e.g., making emotional arguments such as how one's labor will help the speakers of the language and that is compensation enough). One has to set rules and expectations with clear communication on the purpose of a given research project. For instance, when dealing with online communities for data collection and annotation, extra care needs to be shown and benevolent prejudice such as depicting oneself as a savior of a local community (Bird, 2022) must be avoided. Companies and research labs relying on communities for annotation and data creation need to properly compensate the contributors.

The question of **who is annotating what** has to be addressed as well. The scarcity of annotators can lead to poor choices as very often, native speakers cannot be found online easily which has led to researchers choosing people from associated regions—people who do not necessarily speak the language variant in question. This results in a problematic overgeneralisation that puts different languages under the same umbrella simply because they have one or a small set of attributes in common. This often results in potentially oversimplistic solutions. For instance, variations of Arabic differ considerably but numerous research projects have treated entire regions, such as North Africa, as a monolith (e.g., to appear to have more data).

5.2 Be Fair: Give Credit where Credit is Due

Our findings show the unfortunate trend of data workers and NLP practitioners suffering from a lack of recognition, especially those who are part of online communities that focus on low-resource languages. A needed follow-up work would be extensive fieldwork with the various online communities. Hence, our recommendation is a call to action on the setup of fair and comprehensive practices when collaborating with online communities, while taking power differentials into account. That is, existing authorship standards[2] need to be followed and discussed prior to the start of a project as to whether a data worker should be listed as an author. This is particularly critical for junior researchers who substantially contribute to resource

[2] https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html and https://www.aclweb.org/adminwiki/index.php/Authorship_Changes_Policy_for_ACL_Conference_Papers

construction. Moreover, proper financial compensation needs to be provided for annotators who are essential to the construction of large-scale resources. Ideally, a resource paper should provide proof that the annotators were paid and treated fairly if requested by reviewers as recommended by Rogers et al. (2021).

5.3 Choose the Jargon Carefully and Be Aware of False Generalisations

As previously discussed in 5.1, it is important to embrace social awareness and avoid grouping people from colonial and Western perspectives (Bird, 2020, 2022; Held et al., 2023). In this area, we could benefit from critical work in other fields. Hence, one can avoid dismissive and outdated terms and classifications, e.g., “the rest of the world". Note also that *The World’s Values Survey* classification (Haerpfer and Kizilova, 2012), which is often used in NLP papers (e.g., (Santy et al., 2023)), presents an orientalist view of the world (Said, 1977). It has clear flaws such as including Christian-majority countries (e.g., Ethiopia, Rwanda) in a so-called "African-Islamic" category as well as grouping countries that have very little to do with each other (e.g., Kyrgyzstan and Tunisia) leading to misrepresentations.

5.4 Set Fair and Realistic Expectations

As pointed out by (Doğruöz and Sitaram, 2022), tools for low-resource languages are often perceived as scaled-down versions of high-resource ones. Adding to previous work elaborating on what this may mean to the speakers (Bird, 2022; Markl et al., 2024), we focus on the impact of setting these expectations for researchers and practitioners working on mid- to low-resource languages. That is, they may be expected to build models similar to those built for high-resource languages, i.e., tackling the same NLP tasks, and performing extremely well. However, this can be unrealistic for various reasons such as the users’ needs (Blaschke et al., 2024), the language’s specific features (Bird and Yibarbuk, 2024), and the lack of funding linked to the “prestige” of the language as reported by our respondents and similarly discussed by Mihalcea et al. (2024) in the context of LLM research.

No Prescription Joshi et al. (2020) survey the state of NLP in various languages. In fact, people do not necessarily want the tools that researchers think they need. Simultaneously, we should not be limiting what NLP research on mid- to low-resource languages should be about. This is linked to the focus on local communities as this further reinforces the need to communicate with them (Bird and Yibarbuk, 2024; Lent et al., 2022; Mager et al., 2023; Cooper et al., 2024).

Dealing with a "Solved" Problem in a New Language is an Actual Contribution Dealing with what is considered a “solved problem" for high-resource languages does not mean that the research problem is solved for under-served ones–a language may show properties that distinguish it from what is currently available, e.g., a rich morphology or the presence of tones (Adebara and Abdul-Mageed, 2022). Therefore, it is different from what is frequently called “a replication”.

5.5 Check the Source Even if the Language is Low-resource

Due to the limited amount of online data available for mid- to low-resource languages, there is a tendency to use any online sources to build resources for these languages without examining the ethical implications or the appropriateness of the source. While it is typically easier to use religious texts, lyrics, or movie subtitles, these should be carefully considered (Hutchinson, 2024; Mager et al., 2023). For instance, lyrics are not representative of daily communication (Mayer et al., 2008) since, e.g., they often rhyme, and the use of religious texts without a thorough inspection of potential implications can lead to misrepresentations (Mager et al., 2023). Further, we often turn into synthetic data generated using machine translation and LLMs when these show clear limitations, especially in multicultural settings (Hershcovich et al., 2022). It is therefore crucial to investigate what is being translated and to control for the quality of the translation, overgeneralisations, and biases by, e.g., reporting on the performance per each language. Research from other disciplines, even tightly related such as linguistics (Turner, 2023) can help us choose adequate and suitable data sources.

6 Conclusion

We present insights from NLP researchers and practitioners working on under-served languages. We discuss common limitations, research practices in the field, and provide recommendations on how to address the reported issues while remaining fair to data workers. Our work is the first to document NLP researchers and workers’ experiences.

7 Limitations

We acknowledge the fact that there are experiences that are different from those of our respondents and the risk of selection bias. Nonetheless, it is also important to give voice to the concerns of data annotators and researchers working on mid- to low-resource languages, and our survey and this paper aim to do that.

8 Ethical Considerations

While most respondents shared their contact information, it was mainly for following up on the resulting study. We do not share any information that may reveal their identity or the projects they reported.

References

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Steven Bird. 2024. Must nlp be extractive? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14915–14929.

Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8.

Verena Blaschke, Christoph Purschke, Hinrich Schütze, and Barbara Plank. 2024. What do dialect speakers want? a survey of attitudes towards language technology for german dialects. In *Proceedings of ACL*.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Pierre Bourdieu. 1977. The economics of linguistic exchanges. *Social science information*, 16(6):645–668.

Pierre Bourdieu. 1991. Language and symbolic power (ce que parler veut dire). *Polity*.

Karin Knorr Cetina. 1999. *Epistemic cultures: How the sciences make knowledge*. harvard university press.

Ned Cooper, Courtney Heldreth, and Ben Hutchinson. 2024. " it's how you do things that matters": Attending to process to better serve indigenous communities with language technologies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 204–211.

A. Seza Doğruöz and Sunayana Sitaram. 2022. Language technologies for low resource languages: Sociolinguistic and multilingual insights. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 92–97, Marseille, France. European Language Resources Association.

Karën Fort, Gilles Adda, and Kevin Bretonnel Cohen. 2011. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Christian W Haerpfer and Kseniya Kizilova. 2012. The world values survey. *The Wiley-Blackwell Encyclopedia of Globalization*, pages 1–5.

William Held, Camille Harris, Michael Best, and Diyi Yang. 2023. A material lens on coloniality in nlp. *arXiv preprint arXiv:2311.08391*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Julia Hirschberg and Christopher Manning. 2015. Advances in natural language processing. *Science (New York, N.Y.)*, 349:261–266.

Nora Hollenstein, Maria Barrett, and Lisa Beinborn. 2020. Towards best practices for leveraging human language processing signals for natural language processing. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 15–27, Marseille, France. European Language Resources Association.

Ben Hutchinson. 2024. Modeling the sacred: Considerations when using religious texts in natural language processing. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1029–1043, Mexico City, Mexico. Association for Computational Linguistics.

Lilly Irani. 2015. The cultural work of microwork. *New media & society*, 17(5):720–739.

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.

Henry Kahane. 1986. A typology of the prestige language. *Language*, 62(3):495–508.

Bhushan Kotnis, Kiril Gashteovski, Julia Gastinger, Giuseppe Serra, Francesco Alesiani, Timo Sztyler, Ammar Shaker, Na Gong, Carolin Lawrence, and Zhao Xu. 2022. Human-centric research for nlp: Towards a definition and guiding questions. *arXiv preprint arXiv:2207.04447*.

Gavin Leech, Juan J Vazquez, Niclas Kupper, Misha Yagudin, and Laurence Aitchison. 2024. Questionable practices in machine learning. *arXiv preprint arXiv:2407.12220*.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

Manuel Mager, Elisabeth Maier, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897.

Nina Markl, Lauren Hall-Lew, and Catherine Lai. 2024. Language technologies as if people mattered: Centering communities in language technology development. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10085–10099, Torino, Italia. ELRA and ICCL.

Rudolf Mayer, Robert Neumayer, and Andreas Rauber. 2008. Rhyme and style features for musical genre classification by song lyrics. In *Ismir*, volume 14, pages 337–342.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2024. Why ai is weird and should not be this way: Towards ai for everyone, with everyone, by everyone. *arXiv preprint arXiv:2410.16315*.

Saif Mohammad. 2022. Ethics sheets for ai tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8368–8379.

Saif Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'just what do you think you're doing, dave?' a checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Edward W Said. 1977. Orientalism. *The Georgia Review*, 31(1):162–206.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

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

Irina Turner. 2023. Decolonisation through digitalisation? african languages at south african universities. *Curriculum Perspectives*, 43(Suppl 1):73–82.

Heike Winschiers-Theophilus, Shilumbe Chivuno-Kuria, Gereon Koch Kapuire, Nicola J Bidwell, and Edwin Blake. 2010. Being participated: a community approach. In *Proceedings of the 11th Biennial Participatory Design Conference*, pages 1–10.

Diyi Yang, Dirk Hovy, David Jurgens, and Barbara Plank. 2024. The call for socially aware language technologies. *arXiv preprint arXiv:2405.02411*.

A Appendix

A.1 Questionnaire

We would like to investigate the common practices in NLP research on low-resource languages (language variants and "dialects" included).

If you are/were involved in NLP research on low-resource languages, we would like to hear from you. Note that we ****will not**** share your name or demographic information in public. We will only be checking your name for potential follow-up.

(You can also include your initials if you do not want to disclose your name.)

- Email.
- Name.
- *(Optional)* Occupation/Affiliation (if any).
- Which languages do you work on? Language variants and "dialects" included. Please use commas to separate the languages. E.g., language 1, language 2, ...
- What kind of NLP tasks are you interested in? You can name more than one.
- What kind of NLP tools would be relevant/useful for your language(s)?
- Why do you work on this/these language(s) ? You can choose more than one option.
  - I have a genuine interest in languages.
  - I want to build technologies for as many languages as possible.
  - I want to build technologies for my language.
  - Existing technologies in my language of interest suffered from marked limitations.
  - I want to contribute to research on LLMs.
  - I have a genuine interest in NLP/CL/ML.
  - Other. [Note that this is a free text field]
- *(Optional)* If your answer to the previous question included "Existing technologies in my language of interest suffered from marked limitations.", can you tell us why? You can choose more than one option.
  - Resources are scarce.
  - The data is not representative of the language usage.

- The annotation is not performed by fluent speakers.
- The tools do not perform well.
- The tools are not aligned with the needs of the language speakers.
- The tools are not that useful.
- Other. [Note that this is a free text field]

- *(Optional)* If you answered "Existing technologies my language of interest suffered from marked limitations.", can you give an example of these resources or tools?

- *(Optional)* If you answered "Existing technologies my language of interest suffered from marked limitations.", can you share why?

- If you were involved in previous projects, what kind of work were you involved in?
  - Annotation.
  - Data collection.
  - Data creation (e.g., coming up with instructions, questions, etc)
  - Other. [Note that this is a free text field]

- If you were involved in previous projects, did you often get credit for it?
  - Always.
  - Often.
  - Sometimes.
  - Rarely.
  - Never.
  - Other. [Note that this is a free text field]

- *(Optional)* If you were involved in the data collection and/or data annotation in previous projects, how often were you offered authorship?
  - Always.
  - Often.
  - Sometimes.
  - Rarely.
  - Never.
  - Other. [Note that this is a free text field]

- *(Optional)* In projects for which you did not receive financial compensation or authorship, and where you were involved in the data collection and/or data annotation, what was your incentive?
  - I was part of a community.
  - I had access to additional resources (e.g., GPUs, data, etc.).
  - I was acknowledged on the project website.
  - I was acknowledged in the paper.
  - Other. [Note that this is a free text field]

- *(Optional)* For projects where you were simply acknowledged for being an annotator, how long did the data annotation process take?
  - <=2 hours.
  - 2-6 hours.
  - A day of work.
  - 1-7 days.
  - Other. [Note that this is a free text field]

- Are you part of a community? (Yes/No)

- *(Optional)* If you are part of a community, can you name it?

- *(Optional)* Were you involved in projects with industry or academia?
  - Industry.
  - Academia.
  - Both.

- *(Optional long text answer)* Can you name the institutions/projects? (We will not make the names public if you do not want to share the names publicly. See question below.)

- Are you happy making the project names public? (Yes/No)

- *(Optional long text answer)* What are the potential challenges that the NLP/CL community working on the languages that you mentioned face?

- Would you like to receive updates about this project? (Yes/No)

A.2 Languages

The full list of the languages that our respondents have worked is included in the following. Note that participants could work on more than one language.

Named Mid- to Low-resource Languages Afaan Oromo, Albanian, Algerian Arabic, Amharic, Assamese, Awigna, Azerbaijani, Bangla, Basque, Bikol, Cebuano, Coptic, Creole, Croatian, Danish, Egyptian Arabic, Emakhuwa, Faroese, Filipino, Geez, Greek, Harari, Hausa, Hindi, Igbo, Ilocano, Indonesian, Irish, IsiXhosa, Kanuri, Kazakh, Kinyarwanda, Kiswahili, Korean, Light Warlpiri, Lingala, Luganda, Luhya (Lumarachi dialect), Malaysian English, Marathi, Moroccan Arabic, Nepalese, Nyanja, Oromo, Persian/Farsi, Pidgin, Punjabi, Raramuri Russian, Saudi Arabic, Sena, Setswana, Sundanese, Swahili, Tagalog, Tarifit Berber, Tigrinya, Tsonga, Tunisian Arabic, Turkish, Urdu, Warlpiri, Welsh, Wixarika, Wolof, Xhosa, Yoreme Nokki, Yorùbá, Zulu.

Families of Languages African languages, Arabic dialects/variations, English variants, Chatino languages, Gaeilge (including all dialects), Latin American Spanish, Indian languages, Indonesian languages, Nahuatl languages, North African dialects, South East Asian languages.

Named High-resource Languages English, French, Italian, Modern Standard Arabic, Spanish.