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

Nedjma Ousidhoum[1], **Meriem Beloucif**[2], **Saif M. Mohammad**[3]
[1]Cardiff University, [2] Uppsala University, [3] National Research Council Canada
OusidhoumN@cardiff.ac.uk meriem.beloucif@lingfil.uu.se
saif.mohammad@nrc-cnrc.gc.ca

Abstract

Language is a form of symbolic capital that affects people's lives in many ways (Bourdieu, 1977, 1991). As a powerful means of communication, it reflects identities, cultures, traditions, and societies more broadly. Therefore, data in a given language should be regarded as more than just a collection of tokens. Rigorous data collection and labeling practices are essential for developing more human-centered and socially aware technologies. Although there has been growing interest in under-resourced languages within the NLP community, work in this area faces unique challenges, such as data scarcity and limited access to qualified annotators.

In this paper, we collect feedback from individuals directly involved in and impacted by NLP artefacts for medium- and low-resource languages. We conduct both quantitative and qualitative analyses of their responses and highlight key issues related to: (1) data quality, including linguistic and cultural appropriateness; and (2) the ethics of common annotation practices, such as the misuse of participatory research. Based on these findings, we make several recommendations for creating high-quality language artefacts that reflect the cultural milieu of their speakers, while also respecting the dignity and labor of data workers.

1 Introduction

Over the past few years, there has been growing interest in making NLP research more human-centered (Kotnis et al., 2022) and socially aware (Yang et al., 2024). As language technologies are deeply dependent on data quality (Hirschberg and Manning, 2015) and its alignment with the needs of speakers, researchers, and other users, incorporating diverse stakeholder perspectives is essential for building high-quality tools and resources. Traditionally, decisions around data selection, collection, annotation, and model design have been made primarily by researchers. However, the involvement of native speakers, whose languages are at the core of these artefacts, is paramount to better design practices (Bird and Yibarbuk, 2024), since language is inseparable from culture and identity (Bourdieu, 1991). Despite this, when it comes to mid- to low-resource languages, researchers often resort to using any available datasets, seldom scrutinising their quality or relevance due to resource scarcity. While NLP for English and other high-resource languages has benefited from improved standards for corpus quality and ethical research practices (Gebru et al., 2021; Bender and Friedman, 2018; Mohammad, 2022), these standards are not consistently extended to under-resourced languages (Joshi et al., 2020). Therefore, NLP artefacts for underrepresented languages often lack cultural and linguistic grounding, leading to false

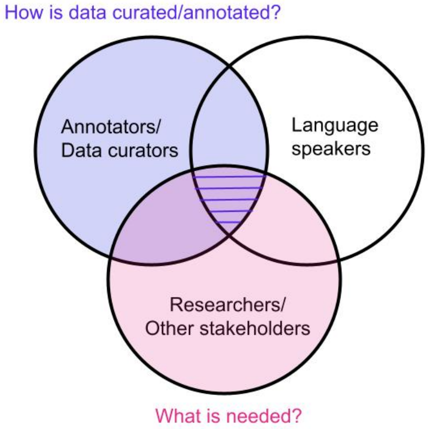
How is data curated/annotated?
Annotators/ Data curators
Language speakers
Researchers/ Other stakeholders
What is needed?

Figure 1: **The main stakeholders and their roles in a data annotation project:** We conducted a survey and reached out to the CL community, specifically NLP researchers and practitioners who have worked on medium- to low-resource languages. Some of the questions focus on the perspectives of a subset of stakeholders highlighted in the figure—namely, speakers who focus on their own languages.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8881–8894
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

generalisations and systemic shortcomings (Bender and Friedman, 2018). This disconnect may hinder meaningful progress and perpetuate inequalities (Blasi et al., 2022; Held et al., 2023), produce sub-optimal user experiences, and reinforce longstanding language hierarchies (Kahane, 1986).

In this position paper, we highlight the current limitations of NLP research for mid- to low-resource languages, specifically in terms of data collection, ethical annotation practices, and overall data quality. We reached out to the CL 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 present the survey's results and provide a set of recommendations based on the responses, focusing on (1) fairness and centering of the language speakers, (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 spans a wide range of topics, including artefact documentation (Bender, 2011; Bender and Friedman, 2018; Gebru et al., 2021; Rogers et al., 2021; Mohammad, 2022) and best practice recommendations (Hollenstein et al., 2020; Mohammad, 2023). Research specifically focused on low-resource languages tends to address the general state of NLP in this area (Held et al., 2023; Joshi et al., 2020; Blasi et al., 2022; Doğruöz and Sitaram, 2022), data collection challenges (Yu et al., 2022), limitations in specific tasks such as machine translation (Mager et al., 2023), developments in LLM research (Mihalcea et al., 2024), and the fundamental issue of including the people whose languages are being studied (Mager et al., 2023; Bird, 2020, 2022; Bird and Yibarbuk, 2024; Lent et al., 2022). Such work highlights the peculiarities of many low-resource languages, the majority of which are vernacular rather than institutionalised or written (Bird and Yibarbuk, 2024; Bird, 2024). It further advocates for language communities to assume agency over their own languages (Schwartz, 2022; Markl et al., 2024; Mihalcea et al., 2024). For instance, Bird and Yibarbuk (2024) examine how experts, such as linguists and computer scientists, collaborate with language communities through participatory design approaches (Winschiers-Theophilus et al., 2010), while Cooper et al. (2024) offer guidance on engaging with Indigenous communities beyond concerns of mere accuracy. Doğruöz and Sitaram (2022) shed light on the need to avoid treating NLP for low-resource languages as a scaled-down version of high-resource language technologies, emphasising the importance of accounting for linguistic and cultural peculiarities. Similarly, Adebara and Abdul-Mageed (2022) stress the significance of such language-specific features (e.g., tones) with a focus on African languages. Beyond language speakers, other work considers the needs of users more broadly. For instance, Blaschke et al. (2024) address the concerns of dialect speakers and emphasise the importance of involving them in the development of language tools and resources. In addition, Yang et al. (2024) define the concept of social awareness and advocate against treating language purely as a computational problem in NLP.

In this paper, we contribute to this ongoing discussion by shifting focus to the practical challenges faced by NLP researchers and practitioners working on mid- and low-resource languages by drawing on methods from social sciences (Cetina, 1999). We investigate the methodological practices and issues currently shaping the field. To the best of our knowledge, little research has examined how NLP work on low-resource languages engages with online communities, apart from a few case studies involving participatory frameworks (e.g., Masakhane) (Birhane et al., 2022), and the work of Lent et al. (2022), who analyse 38 responses from Creole speakers about their experiences with language technologies. Based on an analysis of feedback from our survey respondents, we offer practical recommendations that prioritise transparency and ethically grounded approaches to building more human-centered NLP artefacts for under-served languages.

3 Survey

Our main goal was to investigate the current challenges and problematic practices in NLP research for mid- to low-resource languages and to propose potential solutions. To this end, we reached out to the NLP community (i.e., *CL networks) between June and October 2024 via platforms such as X (formerly Twitter), LinkedIn, Google Groups, Slack channels of online NLP communities, and direct emails. We specifically targeted researchers

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 in both industry and academia.

and annotators working on mid- and low-resource languages, language variants, dialects, and vernaculars, to survey how research in these areas is conducted. Participants reported on common practices, motivations, and key issues. We then conducted both quantitative and qualitative analyses of the responses.

3.1 Respondents

The respondents are NLP researchers and practitioners involved in data collection, annotation, model development, or other research questions related to under-served languages.

3.2 Survey Structure

We ask the respondents about (1) their previous experiences in the area, (2) current problems and limitations related to their language(s) of interest, (3) the motivation behind their involvement in various projects, and (4) how they were credited for tasks often specific to low-resource language research—such as annotation work conducted via online community forums or participatory frameworks.

Note that we allowed respondents to determine for themselves what constitutes a low-resource language, as there is no universally accepted definition. For example, most researchers would agree that Tamasheq is a low-resource language, whereas opinions may differ regarding Malaysian.

3.2.1 General Questions

Respondents had the option to provide their names and contact information for potential follow-up. They were asked about:

- the language(s) they work on,
- the project(s) they have been involved in,
- whether they are or were part of any online community (i.e., participatory research framework),
- whether the project(s) they worked on were based in industry, academia, or both,
- the kinds of NLP tools that are or would be relevant and useful for their language(s) of interest,
- their reasons for working on this/these language(s).

3.2.2 Reporting on Incentives and Potential Limitations

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

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

As we are aware of potential drawbacks in NLP for mid- to low-resource languages (Blasi et al., 2022), we examined whether the respondents had been working 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 any language 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. Moreover, since reaching out to online participatory frameworks is common to projects that focus on under-resourced languages, we asked whether the participants were involved in past projects through such frameworks. This is because involving communities in NLP and ML projects is relatively new to the field and can therefore be abused (purposefully or not) due to the lack of clear standards regard-

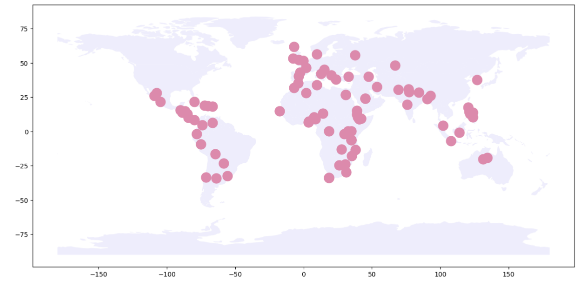

Figure 2: **The main locations** where the languages of interest of our respondents are spoken.

ing data workers in such contexts. Therefore, our questions were the following:

- How often did the respondents receive credit for their contributions? E.g., financial compensation for annotating data.
- 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 credited?

4 Findings

We received 81 responses from researchers working on a wide range of 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 under-resourced languages. Note that percentages do not sum up to 100% as respondents could report on more than one project. That is, participants could also be involved in several languages and NLP tasks. 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

Respondents could report on one or multiple projects they had been involved in. As shown in Table 1, most respondents reported working on academic projects. Around one-third participated in collaborations between industry and academia, or were involved in both types of projects.[1]

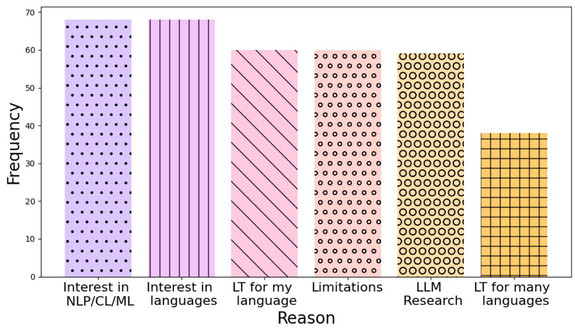

Figure 3: **Frequency of each incentive**. Note that the percentages do not sum up to 100 as the respondents could choose more than one option.

4.1.2 Languages

Among the 81 responses, respondents reported working on over 70 low-resource languages, which they specifically named (see Appendix). Figure 2 illustrates the main regions where these languages are spoken. These include variants, dialects, and vernaculars (e.g., country-specific Arabic dialects), mid- to low-resource languages (e.g., Indonesian), as well as widely acknowledged low-resource languages such as Welsh, Yoreme Nokki, and Setswana. Additionally, around 12% of respondents reported working on language families or branches, such as South Asian languages, all Gaeilige dialects, or Arabic/English varieties. A significant proportion of respondents also work on high-resource languages in parallel.

4.2 Incentives and Potential Limitations

When asking respondents why they had worked on NLP for under-resourced languages, we provided a checklist from which they could select multiple options and add their own entry. We report on common motivations and practices that are typically specific to mid- to low-resource settings, often due to factors such as data scarcity. We also identify problematic instances and analyse the potential reasons behind some of them.

When further examining the common motivations, we report more detailed numbers in Figure 3.

[1] Note that although over 50% of respondents named the projects they participated in, we do not disclose these in order to protect their anonymity.

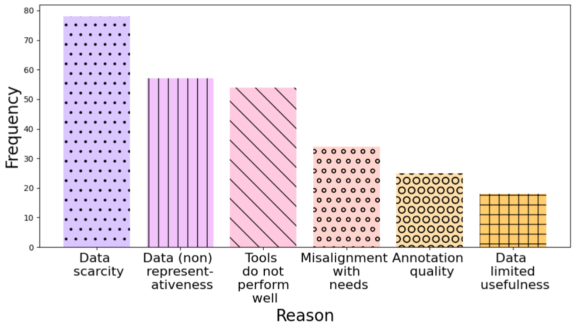

Figure 4: **Frequency of each reported limitation** when the respondents reported working on NLP for low-resource languages due to marked shortcomings.

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. For the respondents whose motivation was to build 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 LLM research (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%); followed by the lack of representativeness of the data (58%) that can manifest in, e.g., unnatural or translationese data instances; the under-performance 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%).

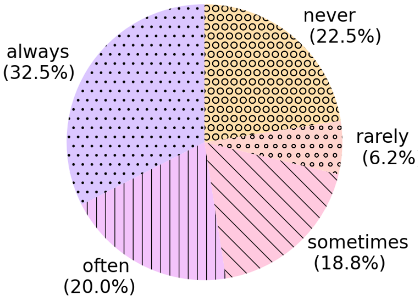

Figure 5: **Respondents on getting credit** for projects they were involved in.

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 they mentioned to justify their choices. When manually processing the answers, we noticed the following recurring topics:

1. **Limitations related to existing resources:** such as their public unavailability, small size, or 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;
   - 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 communities and participatory framework 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. Interestingly, this was equally observed in projects from academia and industry.

We discuss all three of these themes below.

Currently Available Resources Since many under-resourced languages are not institutional but rather vernacular (Bird and Yibarbuk, 2024), collecting data presents considerable challenges when one solely relies on textual data for, 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. It is important to note that what is translated and whether it was further verified by a fluent speaker makes a difference as 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 unnaturalness of the data 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 (i.e., 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. For instance, 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 fluent speakers of mid- to low-resource languages on commonly used annotation platforms such as Amazon Mechanical Turk and Polific. Added to this reason, the lack of funding leads researchers to turn to participatory frameworks. 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, the absence of well-established standards puts vulnerable community members at risk as their efforts may not be properly recognised.

Fundamental Problems Many respondents reported that conducting research in mid- to low-resource languages often entails high costs of data curation and potential outreach to local communities. Further, when resources for an under-served language exist, they are often not freely available.

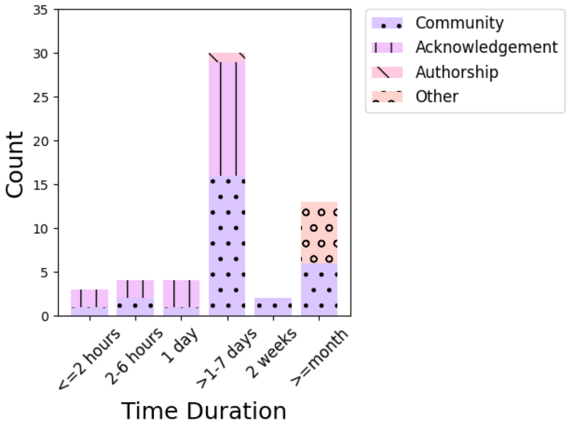

Figure 6: **Respondents on incentives when no proper credit (e.g., financial compensation for data annotation) was offered.** We show the counts of the incentives and the time it took participants to complete their work for a given project (from <=2 hours to more than a month).

4.3 Credit Attribution

We asked the respondents to share whether they were properly credited for their work by, for instance, 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 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 projects in which they were involved, we list the main 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 or a participatory framework (see paragraph below),
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 as we explain below.

The Issue with Over-Reliance on Online Communities and Participatory Research When using standard crowdsourcing platforms such as

AMT or Prolific, researchers can operationalise the annotation process for a given task. Despite their shortcomings (Fort et al., 2011; Irani, 2015), it is possible to protect workers by implementing tests and training for complex annotation tasks. However, for mid- to low-resource languages, platforms such as AMT and Prolific often lack sufficient numbers of registered speakers. As a result, researchers frequently resort to personal networks or participatory approaches. On the one hand, personal connections and community engagement can foster a sense of inclusion and help build trust. On the other hand, however, these approaches carry a significant risk of exploitation and emotional manipulation. That is, some junior researchers were clearly exploited, having been previously told that joining an online community for building language resources was prestigious and worth adding to their CVs. Some respondents expressed frustration at being offered, for instance, company merchandise in exchange for months of labor—a situation that raises further concerns about the involvement of big tech in NLP research for low-resource languages (Abdalla et al., 2023). As shown in Figure 6, 40% of respondents who spent between one day and over a month annotating data reported negative experiences. Their contributions were not adequately compensated, acknowledged, or recognised. This highlights the urgent need for clear guidelines and standards when engaging in community-based or participatory research efforts.

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; Mohammad, 2022, 2023; Leech et al., 2024), our findings substantiate the fact that research on under-resourced languages presents additional challenges linked to the reliance on unconventional practices.

5.1 Center the People

Our findings show that various issues ought to be addressed early as research in the area lacks established standards and is subject to power differentials. For example, some researchers reported that their work and contributions were diminished by more prominent individuals taking first authorship on papers to which they had contributed little or not at all. Many under-resourced 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, 2020; Bird and Yibarbuk, 2024; Bird, 2024; Cooper et al., 2024; Ramponi, 2024). We further reinforce this argument with our findings, i.e., most respondents report marked limitations in their languages (see Figure 4). Hence, when a researcher reaches out to a group of people 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). The question of **who is exactly served** needs to be addressed early to avoid any misconception of perceived needs for language technologies.

Researchers vs. Data Workers In addition to the large proportion of survey respondents who reported not being properly credited for their labor (Figure 5), there were instances of emotional manipulation. This included appeals suggesting that one’s labor would benefit the speakers of the language, and that this alone should be considered sufficient compensation. Notably, such practices were common in participatory research projects initiated by both small groups and Big Tech companies. One has to set rules and expectations with clear communication on the purpose of a given research project. For instance, when dealing with online participatory research 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.

The question of **who is annotating what** has to be addressed as well. The scarcity of qualified annotators can result in poor decision-making. Native speakers are often difficult to find online, which has led some researchers to select individuals from broadly associated regions—people who do not necessarily speak the specific language variant in question. This results in problematic overgeneralisations and overly simplistic solutions, where distinct languages or dialects are grouped together simply because they share one or a few attributes. For example, while Arabic dialects vary significantly, numerous research projects have treated entire regions, such as North Africa, as a linguistic monolith, sometimes to appear to have more data than is truly available.

5.2 Be Fair: Give Credit where Credit is Due

Our findings highlight an unfortunate trend—data workers and NLP researchers, especially those collaborating on participatory research projects, often suffer from a lack of recognition (Figure 3). Accordingly, our recommendation is to set fair and comprehensive practices in participatory research projects, while considering power differentials.

Monetary Compensation Annotators must be provided with proper financial compensation. Companies and research labs that rely on communities for annotation and data creation should ensure fair compensation for contributors, for example, through legally binding contracts. Ideally, resource papers should include evidence that annotators were paid and treated fairly, as recommended by Rogers et al. (2021).

Co-authorship Standards Existing authorship standards must be followed and discussed before the start of a project, particularly with respect to whether data workers should be listed as authors. This is especially important for junior researchers who contribute significantly to resource construction and may expect a leading role in the resulting publication. To inclusively recognise the contributions of junior researchers and those involved in data curation, we recommend the following:

1. Credit all contributors involved in dataset creation, including those that contributed to data curation, design of the annotation setup, and the writing of the paper—as authors.
2. Let data annotators know that they are welcome to play a greater role in the project, possibly rising to the level of co-authors. Discuss with them that meaningful contributions to one or more of the following can lead to co-authorship:
    - Participating in data curation and experimental design.
    - Running language-specific experiments and performing (language-specific) error analyses.
    - Conducting (language-specific) ablation studies.
    - Contributing to the writing of specific paper sections, such as the related work.
    - Providing insights into the resource by:
        - Selecting culturally relevant or representative examples.
        - Offering explanations and interpretations.
        - Describing the annotation process and sharing key observations or challenges.

These tips are adapted from large past annotations efforts that we led (Ousidhoum et al., 2024; Muhammad et al., 2025).

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). 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, 1978). It has clear flaws such as including Christian-majority countries (e.g., Ethiopia, Rwanda) in a so-called “African-Islamic” category and groups of countries that have little in common in one category (e.g., Kyrgyzstan and Tunisia, Bolivia and the Philippines, South Africa and the UAE), leading to misrepresentations and enforcing stereotypes.

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 those developed for high-resource languages. Building on previous work that explores what this may mean for speakers (Bird, 2022; Markl et al., 2024), we shift the focus to researchers and practitioners, who are often expected to build similar models to those created for high-resource languages—i.e., tackling the same NLP tasks and achieving similarly high performance. However, this expectation can be unrealistic for several reasons, including users’ actual needs (Blaschke et al., 2024), the specific linguistic features of the language (Bird and Yibarbuk, 2024), and the lack of funding often tied to the perceived “prestige” of a 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) conducted a survey on the state of NLP across various languages and found that people do not necessarily want the tools that researchers assume they need. One should not prescribe what NLP research on mid- to low-resource languages should be about. The real challenge lies in striking a balance between the technologies that local communities require and the directions pursued by the research community. This balance can be strengthened through ongoing communication and collaboration with various stakeholders (Lent et al., 2022; Mager et al., 2023; Bird and Yibarbuk, 2024; Cooper et al., 2024).

Dealing with a "Solved" Problem in a New Language is an Actual Contribution Just because automatic systems obtain high scores on some English datasets, often making use of massive amounts of English data, does not mean that the problem is solved generally, or that working on that problem in other languages is no longer interesting or valuable. Working with each new language presents many new challenges, e.g., a rich morphology or the presence of tones (Adebara and Abdul-Mageed, 2022). Such work should not be undervalued by calling it "a replication" study.

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

Due to the limited availability of online data for some languages, there is a tendency to use any accessible source to build resources—often without considering the ethical implications or the appropriateness of the content. While it is typically more convenient to rely on religious texts, song lyrics, or film subtitles, such sources should be critically assessed (Hutchinson, 2024; Mager et al., 2023). For example, song lyrics are not representative of everyday communication (Mayer et al., 2008), as they often rhyme and may contain profanity that is not typical of daily language use.

Moreover, scraping Indigenous texts without obtaining proper consent or permissions, and without sensitivity to cultural or religious contexts (Hutchinson, 2024), often results in work on low-resource languages being framed in utilitarian terms, with insufficient attention to deontological concerns. For instance, the use of religious texts without acknowledging their potential implications can lead to cultural misrepresentations, such as portraying certain communities as uniformly religious (Mager et al., 2023).

Additionally, some researchers resort to synthetic data generated via machine translation or large language models (LLMs), despite well-documented limitations, particularly in multicultural or multilingual contexts (Hershcovich et al., 2022) (see Section 4.2.1). Insights from other disciplines, including closely related fields such as linguistics (Turner, 2023), can help guide the selection of more culturally appropriate data sources, especially considering that such datasets are likely to persist and influence future research.

5.6 Position Your Contribution

Similar to Hutchinson (2024), we encourage the inclusion of positionality statements. Specifically, authors should indicate their relationship to the language(s) they work on. This may include, for example, their level of fluency, whether they have formally studied the language(s) in question, and whether they collaborated with native or fluent speakers for tasks such as data annotation. It is also important to clearly state the cultural background of data creators and annotators to avoid false generalisations and regional marginalisation (e.g., treating all Arabic-speaking countries as one homogeneous group (Keleg, 2025)).

6 Conclusion

We present insights from NLP researchers and practitioners working on under-served languages. We discuss reported limitations in this area of research and highlight issues related to data standards, along with common practices and concerning trends, such as problematic incentivisation and the lack of recognition for data workers' labor.
We then offer actionable recommendations to improve data quality by reflecting on data sources, setting realistic expectations for under-resourced languages, and centering speakers and end users while ensuring fairness to data workers.

Limitations

We acknowledge the the risk of selection bias. Nonetheless, our main goal was to give voice to the concerns of data annotators and researchers working on mid- to low-resource languages.

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 identities or the projects they reported on.

Positionality Statement

The three authors are affiliated with governmental and academic institutions in the UK, Sweden, and Canada, respectively. They have experience working with large, diverse groups and in developing new NLP artefacts for high-, mid-, and low-resource languages.

Nedjma Ousidhoum is linguistically proficient in Arabic (Algerian, MSA, and Classical), French, and English. She has lived in Algeria, Hong Kong, and the UK.

Meriem Beloucif is linguistically proficient in Arabic (Algerian and MSA), French, and English, and has a moderate knowledge of German and Swedish. She has lived in Algeria, Hong Kong, Denmark, Germany, and Sweden.

Saif M. Mohammad is proficient in English, Urdu, and Hindi. He has lived in India, USA, and Canada.

7 Acknowledgments

We thank our survey respondents for their insightful comments, and we are grateful to everyone who helped us spread the word on social media or via email.

Many thanks to Joanne Boisson, Michael Schlichtkrull, Sowmya Vajjala, and Evani Radiya-Dixit for their helpful feedback, as well as to the anonymous Action Editor and reviewers for their comments and suggestions.

References

Mohamed Abdalla, Jan Philip Wahle, Terry Ruas, Aurélie Névéol, Fanny Ducel, Saif Mohammad, and Karen Fort. 2023. The elephant in the room: Analyzing the presence of big tech in natural language processing research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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.

Amr Keleg. 2025. LLM alignment for the Arabs: A homogenous culture or diverse ones. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 1–9.

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.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *arXiv preprint arXiv:2502.11926*.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine De Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Winata, Seid Muhie Yimam, and Saif M. Mohammad. 2024. Semrel2024: A collection of semantic textual relatedness datasets for 14 languages.

Alan Ramponi. 2024. Language varieties of Italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

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. 1978. *Orientalism*. Pantheon Books, New York.

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*.

Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. 2022. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3725–3743.

Appendix

A 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 them 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)

B 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. Some may have also conducted research for high-resource languages and the respondents may or may not speak the language(s).

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.