Building Representative Corpora from Illiterate Communities: A Review of Challenges and Mitigation Strategies for Developing Countries

Stephanie Hirmer[1], **Alycia Leonard**[1], **Josephine Tumwesige**[2], **Costanza Conforti**[2,3]
[1]Energy and Power Group, University of Oxford
[2]Rural Senses Ltd.
[3]Language Technology Lab, University of Cambridge
stephanie.hirmer@eng.ox.ac.uk

Abstract

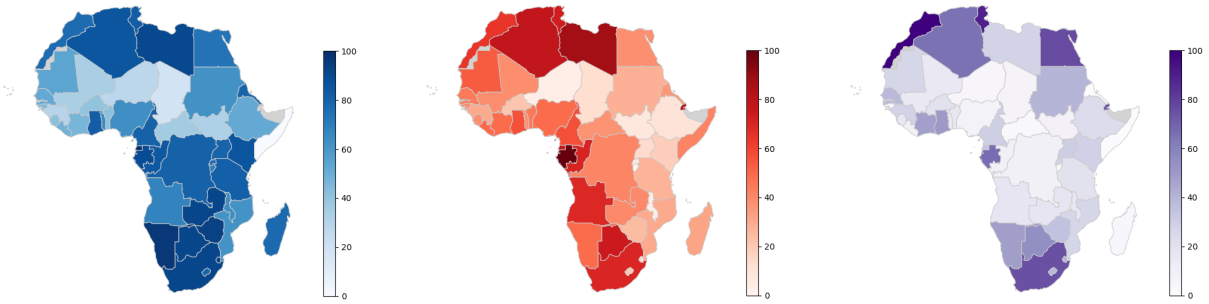
Most well-established data collection methods currently adopted in NLP depend on the assumption of speaker literacy. Consequently, the collected corpora largely fail to represent swathes of the global population, which tend to be some of the most vulnerable and marginalised people in society, and often live in rural developing areas. Such underrepresented groups are thus not only ignored when making modeling and system design decisions, but also prevented from benefiting from development outcomes achieved through data-driven NLP. This paper aims to address the under-representation of illiterate communities in NLP corpora: we identify potential biases and ethical issues that might arise when collecting data from rural communities with high illiteracy rates in Low-Income Countries, and propose a set of practical mitigation strategies to help future work.

1 Introduction

The exponentially increasing popularity of supervised Machine Learning (ML) in the past decade has made the availability of data crucial to the development of the Natural Language Processing (NLP) field. As a result, much NLP research has focused on developing rigorous processes for collecting large corpora suitable for training ML systems. We observe, however, that many best practices for quality data collection make two implicit assumptions: that speakers have internet access and that they are literate (i.e. able to read and often write text effortlessly[1]). Such assumptions might be reasonable in the context of most High-Income Countries (HICs) (UNESCO, 2018). However, in Low-Income Countries (LICs), and especially in sub-Saharan Africa (SSA), such assumptions may not hold, particularly in rural developing areas where the bulk of the population lives (Roser and Ortiz-Ospina (2016), Figure 1). As a consequence, common data collection techniques – designed for use in HICs – fail to capture data from a vast portion of the population when applied to LICs. Such techniques include, for example, crowdsourcing (Packham, 2016), scraping social media (Le et al., 2016) or other websites (Roy et al., 2020), collecting articles from local newspapers (Marivate et al., 2020), or interviewing experts from international organizations (Friedman et al., 2017). While these techniques are important to easily build large corpora, they implicitly rely on the above-mentioned assumptions (i.e. internet access and literacy), and might result in demographic misrepresentation (Hovy and Spruit, 2016). In this paper, we make a first step towards addressing *how to build representative corpora in LICs from illiterate speakers*. We believe that this is a currently unaddressed topic within NLP research. It aligns with previous work investigating sources of bias resulting from the under-representation of specific demographic groups in NLP corpora (such as women (Hovy, 2015), youth (Hovy and Søgaard, 2015), or ethnic minorities (Groenwold et al., 2020)). In this paper, we make the following contributions: (i) we introduce the challenges of collecting data from illiterate speakers in §2; (ii) we define various possible sources of biases and ethical issues which can contribute to low data quality we define various possible sources of biases and ethical issues which can contribute to low data quality we define various possible sources of biases and ethical issues which can contribute to low data quality §3; finally, (iii) drawing on years of experience in data collection in LICs, we outline practical countermeasures to address these issues in §4.

[1]For example, input from speakers is often taken in writing, in response to a written stimulus which must be read.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2176–2189
April 19 - 23, 2021.



(a) Adult literacy (% ages 15+, UNESCO (2018))

(b) Urban population (% total, UNDESA (2018))

(c) Internet usage (% of total, ITU (2019))

Figure 1: Literacy, urban population, and internet usage in African countries. Note that countries with more rural populations tend to have less literacy and less internet users. These countries are likely to be under-represented in corpora generated using common data collection methods that assume literacy and internet access (Grey: no data).

2 Listening to the Illiterate: What Makes it Challenging?

In recent years, developing corpora that encompasses as many human languages as possible has been recognised as important in the NLP community. In this context, widely translated texts (such as the Bible (Mueller et al., 2020) or the Human Rights declaration (King, 2015)) are often used as a source of data. However, these texts tend to be quite short and domain-specific. Moreover, while the Internet constitutes a powerful data collection tool which is more representative of real language use than the previously-mentioned texts, it excludes illiterate communities, as well as speakers which lack reliable internet access (as is often the case in rural developing settings, Figure 1).

Given the obstacles to using these common language data collection methods in LIC contexts, the NLP community can learn from methodologies adopted in other fields. Researchers from fields such as sustainable development (SD, Gleitsmann et al. (2007)), African studies (Adams, 2014), and ethnology (Skinner et al., 2013), tend to rely heavily on qualitative data from oral interviews, transcribed verbatim. Collecting such data in rural developing areas is considerably more difficult than in developed or urban contexts. In addition to high illiteracy levels, researchers face challenges such as seasonal roads and low population densities. To our knowledge, there are very few NLP works which explicitly focus on building corpora from rural and illiterate communities: of those works that exist, some present clear priming effect issues (Abraham et al., 2020), while others focus on application (Conforti et al., 2020). A detailed description of best practices for data collection remains a notable research gap.

3 Definitions and Challenges

Guided by research in medicine (Pannucci and Wilkins, 2010), sociology (Berk, 1983), and psychology (Gilovich et al., 2002), NLP has experienced increasing interest in ethics and bias mitigation to minimise unintentional demographic misrepresentation and harm (Hovy and Spruit, 2016). While there are many stages where bias may enter the NLP pipeline (Shah et al., 2019), we focus on those pertinent to *data collection* from rural illiterate communities in LICs, leaving the study of biases in model development for future work[2].

3.1 Data Collection Biases

Biases in data collection are inevitable (Marshall, 1996) but can be minimised when known to the researcher (Trembley, 1957). We identify various biases that can emerge when collecting language data in rural developing contexts, which fall under three broad categories: sampling, observer, and response bias. Sampling determines who is studied, the interviewer (or observer) determines what information is sought and how it is interpreted, and the interviewee (or respondent) determines which information is revealed (Woodhouse, 1998). These categories span the entire data collection process and can affect the quality and quantity of language data obtained.

[2]Note, this paper does not focus on a particular NLP application, as once the data has been collected from illiterate communities it can be annotated for virtually any specific task.

3.2 Sampling or selection bias

Sampling bias occurs when observations are drawn from an unrepresentative subset of the population being studied (Marshall, 1996) and applied more widely. In our context, this might arise when selecting communities from which to collect language data, or specific individuals within each community. When sampling communities, bias can be introduced if convenience is prioritized. Communities which are easier to access may not produce language data representative of a larger area or group. This can be illustrated through Uganda's refugee response, which consists of 13 settlements (including the 2nd largest in the world) hosted in 12 districts (UNHCR, 2020). Data collection may be easier in one of the older, established settlements; however, such data cannot be generalised over the entire refugee response due to different cultural backgrounds, length of stay of refugees in different areas, and the varied stages along the humanitarian chain – emergency, recovery or development – found therein (Winter, 1983; OECD, 2019). Prioritizing convenience in this case may result in corpora which over-represents the cultural and economic contexts of more established, longer-term refugees. When sampling interviewees, bias can be introduced when certain sub-sets of a community have more data collected than others (Bryman, 2012). This is seen when data is collected only from men in a community due to cultural norms (Nadal, 2017), or only from wealthier people in cell-phone-based surveys (Labrique et al., 2017).

3.2.1 Observer bias

Observer bias occurs when there are systematic errors in how data is recorded, which may stem from observer viewpoints and predispositions (Gonsamo and D'Odorico, 2014). We identify three key observer biases relevant to our context.

Firstly, **confirmation bias**, which refers to the tendency to look for information which confirms one's preconceptions or hypotheses (Nickerson, 1998). Researchers collecting data in LICs may expect interviewees to express needs or hardships based on their preconceptions. As Kumar (1987) points out, "often they hear what they want to hear and ignore what they do not want to hear". A team conducting a needs assessment for a rural electrification project, for instance, may expect a need for electricity, and thus consciously or subconsciously seek data which confirms this, interpret potentially unrelated data as electricity-motivated (Hirmer and Guthrie, 2017), or omit data which contradicts their hypothesis (Peters, 2020). Using such data to train NLP models may introduce unintentional bias towards the original expectations of the researchers instead of accurately representing the community.

Secondly, the interviewer's understanding and interpretation of the speaker's utterances might be influenced by their class, culture and language. Note that, particularly in countries without strong language standardisation policies, consistent semantic shifts can happen even between varieties spoken in neighboring regions (Gordon, 2019), which may result in systematic **misunderstanding** (Sayer, 2013). For example, in the neighboring Ugandan tribes of Toro and Bunyoro, the same word *omunyoro* means respectively *husband* and *a member of the tribe*. Language data collected in such contexts, if not properly handled, may contain inaccuracies which lead to NLP models that misrepresent these tribes. Rich information communicated through gesture, expression, and tone (i.e. nonverbal data, Oliver et al. (2005)) may also be systematically lost during verbatim transcription, causing inadvertent inconsistencies in the corpora.

Thirdly, **interviewer bias**, which refers to the subjectivity unconsciously introduced into data gathering by the worldview of the interviewer (Frey, 2018). For instance, a deeply religious interviewer may unintentionally frame questions through religious language (e.g. *it is God's will*, *thank God*, etc.), or may perceive certain emotions (e.g. thankfulness) as inherently religious, and record language data including this perception. The researcher's attitude and behaviour may also influence responses (Silverman, 2013); for instance, when interviewers take longer to deliver questions, interviewees tend to provide longer responses (Matarazzo et al., 1963). Unlike in internet-based language data collection, where all speakers are exposed to uniform, text-based interfaces, collecting data from illiterate communities necessitates the presence of an interviewer, who cannot always be the same person due to scalability constraints, introducing this inevitable variability and subsequent data bias.

3.2.2 Response bias

Response bias occurs when speakers provide inaccurate or false responses to questions. This is particularly important when working in rural settings, where the majority of data collection is currently

related to SD projects. The majority of existing data is biased by the projects for which it has been collected, and any newly collected data for NLP uses is also likely to be used in decision making for SD. This inherent link of data collection to material development outcomes inevitably affects what is communicated. There are five key response biases relevant to our context.

Firstly, **recall bias**, where speakers recall only certain events or omit details (Coughlin, 1990). This is often as a result of external influences, such as the presence of a data collector who is new to the community. Recall can also be affected by the distortion or amplification of traumatic memories (Strange and Takarangi, 2015); if data is collected around a topic a speaker may find traumatic, recall bias may be unintentionally introduced.

Secondly, **social desirability bias**, which refers to the tendency of interviewees to provide socially desirable/acceptable responses rather than honest responses, particularly in certain interview contexts (Bergen and Labonté, 2020). In tight-knit rural communities, it may be difficult to deviate from traditional social norms, leading to biased data. As an illustrative example, researchers in Nepal found that interviewer gender affected the detail in responses to some sensitive questions (e.g. sex and contraception): participants provided less detail to male interviewers (Axinn, 1991). Social desirability bias can produce corpora which misrepresent community social dynamics or under-represent sensitive topics.

Thirdly, **recency effect or serial-position**, which is the tendency of a person to recall the first and last items in a series best, and the middle items worst (Troyer, 2011). This can greatly impact the content of language data. For instance, in the context of data collection to guide development work, it is important to understand current needs and values (Hirmer and Guthrie, 2016); however, if only the most recent needs are discussed, long-term needs may be overlooked. To illustrate, while a community which has just experienced a poor agricultural season may tend to express the importance of improving agricultural output, other needs which are less top-of-mind (i.e. healthcare, education) may be equally important despite being expressed less frequently. If data containing *recency bias* is used to develop NLP models, particularly for sustainable development applications (such as for Automatic UPV Classification, Conforti et al. (2020)), these may amplify current needs and under-represent long-term needs.

Fourthly, **acquiescence bias**, also known as "yea" saying (Laajaj and Macours, 2017), which can occur in rural developing contexts when interviewees perceive that certain (possibly false) responses will please a data collector and bring benefits to their community. For example, if data collection is being undertaken by a group with a stated desire to build a school may be more likely to hear about how much education is valued.

Finally, **priming effect**, or the ability of a presented stimulus to influence one's response to a subsequent stimulus (Lavrakas, 2008). Priming is problematic in data collection to inform SD projects; it can be difficult to collect data on the relative importance of simultaneous (or conflicting) needs if the community is primed to focus on one (Veltkamp et al., 2011). An example is shown in Figure 2a; respondents may be drawn to speak more about the most dominant prompts presented in the chart. This is typical of a broader failure in SD to uncover beneficiary priorities without introducing project bias (Watkins et al., 2012). Needs assessments, like the one referenced above linked to a rural electrification project, tend to focus explicitly on project-related needs instead of more broadly identifying what may be most important to communities (Masangwi, 2015; USAID, 2006). As speakers will usually know why data is being collected in such cases, they may be biased towards stating the project aim as a need, thereby skewing the corpora to over-represent this aim.

3.3 Ethical Considerations

Certain ethical codes of conduct must be followed when collecting data from illiterate speakers in rural communities in LICs (Musoke et al., 2020). Unethical data collection may harm communities, treat them without dignity, disrupt their lives, damage intra-community or external relationships, and disregard community norms (Thorley and Henrion, 2019). This is particularly critical in rural developing regions, as these areas are home to some of the world's poorest and most vulnerable to exploitation (Christiaensen and Subbarao, 2005; de Cenival M., 2008). Unethical data collection can replicate extractive colonial relationships whereby data is extracted from communities with no mutual benefit or ownership (Dunbar and Scrimgeour, 2006). It can lead to a lack of trust between data collec-

tor and interviewees and unwillingness to participate in future research (Clark, 2008). These phenomena can bias data or reduce data availability. Ethical data collection practices in rural developing regions with high illiteracy include: obtaining consent (McAdam, 2004), accounting for cultural differences (Silverman, 2013), ensuring anonymity and confidentiality (Bryman, 2012), respecting existing community or leadership structures (Harding et al., 2012), and making the community the owner of the data. While the latter is not often currently practiced, it is an important consideration for community empowerment, with indigenous data sovereignty efforts (Rainie et al., 2019) already setting precedent.

4 Countermeasures

Drawing on existing literature and years of field experience collecting spoken data in LICs, below we outline a number of practical data collection strategies to minimise previously-outlined challenges (§3), enabling the collection of high-quality, minimally-biased data from illiterate speakers in LICs suitable for use in NLP models. While these measures have primarily been applied in SSA, we have also successfully tested them in projects focusing on refugees in the Middle East and rural communities in South Asia.

4.1 Preparation

Here, we outline practical preparation steps for careful planning, which can minimise error and reduce fieldwork duration (Tukey, 1980).

Local Context. A thorough understanding of local context is key to successful data collection (Hentschel, 1999; Bukenya et al., 2012; Launiala and Kulmala, 2006). Local context is broadly defined as facts, concepts, beliefs, values, and perceptions used by local people to interpret the world around them, and is shaped by their surroundings (i.e. their worldview, Vasconcellos and Vasconcellos Sobrinho (2014)). It is important to consider local context when preparing to collect data in rural developing areas, as common data collection methods may be inappropriate due to contextual linguistic differences and deep-rooted social and cultural norms (Walker and Hamilton, 2011; Mafuta et al., 2016; Nikulina et al., 2019; Wang et al.). Selecting a contextually-appropriate data collection method is critical in mitigating *social desirability bias* in the collected data, among other challenges. Researchers should review socio-economic surveys and/or consult local stakeholders who can offer valuable insights on practices and social norms. These stakeholders can also highlight current or historical matters of concern to the area, which may be unfamiliar to researchers, and reveal local, traditional, and indigenous knowledge which may impact the data being collected (Wu, 2014) and result in *recency effect*. It is good practice to identify local conflicts and segmentation within a community, especially in a rural context, where the population is vulnerable and systematically unheard (Dudwick et al., 2006; Mallick et al., 2011).

Case sampling. In qualitative research, sample cases are often strategically selected based on the research question (i.e. *systematic* or *purposive* sampling, Bryman (2012)), and characteristics or circumstances relevant to the topic of study (Yach, 1992). If data collected in such research is used beyond its original scope, *sampling bias* may result. So, while data collected in previous research should be re-used to expand NLP corpora where possible, it is important to be cognizant of the purposive sampling underlying existing data. A comprehensive dataset characterisation (Bender and Friedman, 2018; Gebru et al., 2018) can help researchers understand whether an existing dataset is appropriate to use in new or different research, such as in training new NLP models, and can highlight the potential ethical concerns of data re-use.

Participant sampling. Interviewees should be selected to represent the diverse interests of a community or sampling group (e.g. occupation, age, gender, religion, ethnicity or male/female household heads (Bryman, 2012)) to reduce *sampling bias* (Kitzinger, 1994). To ensure representativity in collected data, sampling should be random, i.e. every subject has equal probability to be included (Etikan et al., 2016). There may be certain societal subsets that are concealed from view (e.g. as a result of embarrassment from disabilities or physical differences) based on cultural norms in less inclusive societies (Vesper, 2019); particular care should be exercised to ensure such subsets are represented.

Group composition. Participant sampling best practices vary by data collection method, with particular care being necessary in group settings. In traditional societies where strong power dynamics exist, attention should be paid to group composition and interaction to prevent some voices from

	Bias & Definition	Key countermeasures
Sampling	**Community**: An unrepresentative sample set is generalised over the entire case being studied. **Participant**: Certain sub-sets of a community have more data collected from them than others.	• Select representative communities & only apply data within same scope (i.e. consult data statements) • Select representative participants, only apply data within same scope & avoid tempting rewards
Observer	**Confirmation**: Looking for information that confirms one's preconceptions or hypotheses about a topic/research/sector. **Misunderstanding**: Data is incorrectly transcribed or categorized as a result of class, cultural, or linguistic differences. **Interviewer**: Unconscious subjectivity introduced into data gathering by interviewers' worldview.	• Employ interviewers that are impartial to the topic/research/sector investigated. • Employ local people & minimise # of people involved for both data collection & transcription. • Undertake training to minimise influence exerted from questions, technology, & attitudes.
Response	**Recall**: Tendency of speakers to recall only certain events or omit details **Social-desirability**: Tendency of participants to provide socially desirable/acceptable responses rather than to respond honestly. **Recency effect**: Tendency to recall first or last items in a series best, & middle items worst. **Acquiescence**: Respondents perceive certain, perhaps false, answers may please data collectors, bringing community benefits. **Priming effect**: Ability of a presented stimulus to influence one's response to a subsequent stimulus	• Collect support data (e.g. from socio-economic data or local stakeholders) to compare with interviews. • Select interviewers & design interview processes to account for known norms which might skew responses • Minimise external influence on participants throughout data gathering (e.g. technologies, people, perceptions). • Gather non-sectoral holistic insights (e.g. from socio-economic data or local stakeholders) • Use appropriate visual prompts (graphically similar), language and technology

Table 1: Sources of potential bias in data collection when operating in rural and illiterate settings in developing countries, and key countermeasures that can help mitigating them.

being silenced or over-represented (Stewart et al., 2007). For example, in Uganda, female interviewees may be less likely to voice opinions in the presence of male interviewees (FIDH, 2012; Axinn, 1991), introducing a form of *social desirability bias* in resulting corpora. To minimise this risk of data bias, relations and power dynamics must be considered during data collection planning (Hirmer, 2018). It may be necessary to exclude, for instance, close relatives, governmental officials, and village leaders from group discussions where data is being collected, and instead engage such stakeholders in separate activities to ensure that their voices are included in the corpora without biasing the data collected from others.

Interviewer selection. The interviewer has a significant opportunity to introduce *observer and response biases* in collected data (Salazar, 1990). Interviewers familiar with local language, including community-specific dialects, should be selected wherever possible. Moreover, to reduce *misunderstanding* and *recall biases* in collected data, it is useful to have the same person who conducts the interviews also transcribe them. This minimizes the layers of linguistic interpretation affecting the final dataset and can increase accuracy through familiarity with the interview content. If the interviewer is unavailable, the transcriber must be properly trained and briefed on the interviews, and made aware of the level of detail needed during transcription (Parcell and Rafferty, 2017).

Study design. In rural LIC communities, qualitative data like natural language is usually collected by observation, interview, and/or focus group discussion (or a combination, known as mixed methods) which are transcribed verbatim (Moser and Korstjens, 2018). Prompts are often used to spark discussion. Whether visual prompts (Hirmer, 2018) or verbalised question prompts are used during data collection, these should be designed to: (i) accommodate illiteracy, (ii) account for disabilities (e.g. visually impairment; both could cause *sampling bias*), and (iii) minimise bias towards a topic or sector (e.g. minimising *acquisition bias* and *confirmation bias*). For instance, visual prompts should be graphically similar and contain only visuals familiar to the respondents. This is analogous to the uniform interface with which speakers interact during text-based online data collection, where the platform used is graphically the same to all users inputting data. Using varied graphical styles or unfamiliar images may result in *priming* (Figure 2a). To minimise *recall bias* or *recency effect* in collected data, socio-economic data can be integrated in data analysis to better understand if the assertions made in collected data reference recent events, for example. These should be non-sector specific, to gain holistic insights and to minimise

acquisition bias and *confirmation bias*.

4.2 Engagement

Here, we outline practical steps for successful community engagement to achieve ethical and high-quality data collection.

Defining community. Defining a community in an open and participatory manner is critical to meaningful engagement (Dyer et al., 2014). By understanding the community the way they understand themselves, misunderstandings and tensions that affect data quality can be minimized. The definition of the community (MacQueen et al., 2001) coupled with the requirements and use-cases for the collected data determines the data collection methodology and style which will be most appropriate (e.g. interview-based community consultation vs. collaborative co-design for mutual learning).

Follow formal structures. Researchers entering a community where they have no background to collect data should endeavour to know the community prior to commencing any work (Diallo et al., 2005). This could entail visiting the community and mapping its hierarchies of authority and decision-making pathways, which can guide the research team on how to interact respectfully with the community (Tindana et al., 2011). This process should also illuminate whether knowledgeable community members should facilitate entry by performing introductions and assisting the external data collection team. Following formal community structures is vital, especially in developing communities, where traditional rules and social conventions are strongly held yet often not articulated explicitly or documented. Approaching community leaders in the traditional way can help to build a positive long-term relationship, removing suspicion about the nature and motivation of the researchers' activities, explaining their presence in the community, and most importantly building trust as they are granted permission to engage the community by its leadership (Tindana et al., 2007).

Verbalising consent. Data ethics is paramount for research involving human participants (Accenture, 2016; Tindana et al., 2007), including any collection of personal and identifiable data, such as natural language. Genuine (i.e. voluntary and informed) consent must be obtained from interviewees to prevent use of data which is illegal, coercive, or for a purpose other than that which has been agreed (McAdam, 2004). The Nuffield Council on Bioethics (2002) caution that in LICs, misunderstandings may occur due to cultural differences, lower social-economic status, and illiteracy (McMillan et al., 2004) which can call into question the legitimacy of consent obtained. Researchers must understand that methods such as long information forms and consent forms which must be signed may be inappropriate for the cultural context of LICs and can be more likely to confuse than to inform (Tekola et al., 2009). The authors advise that consent forms should be verbal instead of written, with wording familiar to the interviewees and appropriate to their level of comprehension (Tekola et al., 2009). For example, to speak of data storage on a password protected computer while obtaining consent in a rural community without access to electricity or information technology is unfitting. Innovative ways to record consent can be employed in such contexts (e.g. video taping or recording), as signing an official document may be "viewed with suspicion or even outright hostility" (Upjohn and Wells, 2016), or seen as "committing ... to something other than answering questions". Researchers new to qualitative data collection should seek advice from experienced researchers and approval from their ethics committee before implementing consent processes.

Approaching participants. Despite having gained permission from community authorities and obtained consent to collect data, researchers must be cautious when approaching participants (Irabor and Omonzejele, 2009; Diallo et al., 2005) to ensure they do not violate cultural norms. For example, in some cultures a senior family member must be present for another household member to be interviewed, or a female must be accompanied by a male counterpart during data collection. Insensitivity to such norms may compromise the data collection process; so, they should be carefully noted when researching local context (§4.1) and interviews should be designed to accommodate them where possible. Furthermore, researchers should investigate the motivations of the participants to identify when inducements become inappropriate and may lead to either harm or data bias (McAdam, 2004).

Minimise external influence. Researchers must be aware of how external influences can affect data collection (Ramakrishnan et al., 2012). We find three main levels of external influence: (i) technologies unfamiliar to a rural developing country

context may induce *social desirability bias* or *priming* (e.g. if a researcher arrives to a community in an expensive vehicle or uses a tablet for data collection); (ii) intergroup context, which according to Abrams (2010) refers to when "people in different social groups view members of other groups" and may feel prejudiced or threatened by these differences. This can occur, for instance, when a newcomer arrives and speaks loudly relative to the indigenous community, which may be perceived as overpowering; (iii) there is the risk of a researcher over-incentivizing the data collection process, using leading questions and judgemental framing (*interviewer bias* or *confirmation bias*). To overcome these influences, researchers must be cognizant of their influence and minimise it by hiring local mediators where possible alongside employing appropriate technology, mannerisms, and language.

4.3 Undertaking Interviews

Here, we detail practical steps to minimise challenges during the actual data collection.

Interview settings. People have personal values and drivers that may change in specific settings. For example, in the Ugandan Buganda and Busoga tribes, it is culturally appropriate for the male head if present to speak on behalf of his wife and children. This could lead to corpora where input from the husband is over-represented compared to the rest of the family. To account for this, it is important to collect data in multiple interview settings (e.g. individual, group male/female/mixed; Figures 2b, 2c). Additionally, the inputs of individuals in group settings should be considered independently to ensure all participants have an equal say, regardless of their position within the group (Barry et al., 2008; Gallagher et al., 1993). This helps to avoid *social desirability bias* in the data and is particularly important in various developing contexts where stereotypical gender roles are prominent (Hirmer, 2018). During interviews, verbal information can be supplemented through the observation of tone, cadence, gestures, and facial expressions (Narayanasamy, 2009; Hess et al., 2009), which could enrich the collected data with an additional layer of annotation.

Working with multiple interviewers. Arguably, one of the biggest challenges in data collection is ensuring consistency when working with multiple interviewers. Some may report word-for-word what is being said, while others may summarise or misreport, resulting in systematic *misunderstanding*. Despite these risks, employing multiple interviewers is often unavoidable when collecting data in rural areas of developing countries, where languages often exhibit a high number of regional, non-mutually intelligible varieties. This is particularly prominent across SSA. For example, 41 languages are spoken in Uganda (Nakayiza, 2016); English, the official language, is fluently spoken by only ∼5% of the population, despite being widely used among researchers and NGOs (Katushemererwe and Nerbonne, 2015). To minimise data inconsistency, researchers should: (i) undertake interviewer training workshops to communicate data requirements and practice data collection processes through mock field interviews; (ii) pilot the data collection process and seek feedback to spot early deviation from data requirements; (iii) regularly spot-check interview notes; (iv) support written notes with audio recordings[3]; and (v) offer quality based incentives to data collectors.

Participant remuneration. While it is common to offer interviewees some form of remuneration for their time, the decision surrounding payment is ethically-charged and widely contested (Hammett and Sporton, 2012). Rewards may tempt people to participate in data collection against their judgement. They can introduce *sampling bias* or create power dynamics resulting in *acquiescence bias* (Largent and Lynch, 2017). Barbour (2013) offers three practical solutions: (i) not advertise payment; (ii) omit the amount being offered; or (iii) offer non-financial incentives (e.g. products that are desirable but difficult to get in an area). The decision whether or not to remunerate should not be based upon the researcher's own ethical beliefs and resources, but instead by considering the specific context[4], interviewee expectations, precedents set by previous researchers, and local norms (Hammett and Sporton, 2012). Representatives from local organisations (such as NGOs or governmental authorities) may be able to offer advice.

[2]While participants' photographing permission was granted, photos were pixelised to protect identity.

[3]Relying only on audio data recording may be risky: equipment can fail or run out of battery (which is not easily remedied in rural off-grid regions) and seasonal factors (as noise from rain on corrugated iron sheets, commonly used for roofing in SSA) can make recordings inaudible (Hirmer, 2018)).

[4]In rural Uganda, for example, politicians commonly engage in *vote buying* by distributing gifts (Blattman et al., 2019) such as soap or alcohol. It is therefore considered an unruly form of remuneration and can only be avoided when known.

(a) (b) (c)

Figure 2: Collecting oral data in rural Uganda. 2a *Priming effect* (note the word "Energy" in the poster's title and the visual prompts differences between items). On the contrary, 2b and 2c show minimal priming; note also that different demographics are separately interviewed (women group, single men) to avoid *social desirability bias*.

4.4 Post-interviewing

Here, we discuss practical strategies to mitigate ethical issues surrounding the management and stewardship of collected data.

Anonymisation. To protect the participants' identity and data privacy, locations, proper names, and culturally explicit aspects (such as tribe names) of collected data should be made anonymous (Sweeney, 2000; Kirilova and Karcher, 2017). This is particularly important in countries with security issues and low levels of democracy.

Safeguarding data. A primary responsibility of the researcher is to safeguard participants' data (Kirilova and Karcher, 2017). In addition to anonymizing data, mechanisms for data management include in-place handling and storage of data (UKRI, 2020a). Whatever data management plan is adopted, it must be clearly articulated to participants before the start of the interview (i.e. as part of the consent process (Silverman, 2013)), as was discussed in §4.2 (*Verbalising consent*).

Withdrawing consent. Participants should have the ability to withdraw from research within a specified time frame. This is known as *withdraw consent* and is commonly done by phone or email (UKRI, 2020b). As people in rural illiterate communities have limited means and technology access, a local phone number and contact details of a responsible person in the area should be provided to facilitate withdraw consent.

Communication and research fatigue. While researchers frequently extract knowledge and data from communities, only rarely are findings fed back to communities in a way that can be useful to them. Whatever the research outcomes, researchers should share the results with participating communities in an appropriate manner. In illiterate communities, for instance, murals (Jimenez, 2020), artwork, speeches, or song could be used to communicate findings. Not communicating findings may result in research fatigue as people in *over-studied* communities are no longer willing to participate in data collection. This is common "where repeated engagements do not lead to any experience of change [...]" Clark (2008). Patel et al. (2020) offers practical guidance to minimise research fatigue by: (i) increasing transparency of research purpose at the beginning of the research, and (ii) engaging with gatekeeper or oversight bodies to minimise number of engagements per participant. Failure to restrict the number of times that people are asked to participate in studies risks poor future participation (Patel et al., 2020) which can also lead to *sampling bias*.

5 Conclusion

In this paper, we provided a first step towards defining best practices in data collection in rural and illiterate communities in Low-Income Countries to create globally representative corpora. We proposed a comprehensive classification of sources of bias and unethical practices that might arise in the data collection process, and discussed practical steps to minimise their negative effects. We hope that this work will motivate NLP practitioners to include input from rural illiterate communities in their research, and facilitate smooth and respectful interaction with communities during data collection. Importantly, despite the challenges that working in such contexts might bring, the effort to build substantial and high-quality corpora which represent this subset of the population can result in considerable SD outcomes.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. We are also grateful to Claire McAlpine, as well as Malcolm McCulloch and other members of the Energy and Power Group (University of Oxford) for providing valuable feedback on early versions of this paper. This research was carried out as part of the Oxford Martin Programme on Integrating Renewable Energy. Finally, we are grateful to the Rural Senses team for sharing experiences on data collection.

References

Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing speech data for low-resource languages from low-income workers. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2819–2826, Marseille, France. European Language Resources Association.

Dominic Abrams. 2010. *Processes of prejudices: Theory, evidence and intervention*. Equalities and Human Rights Commission.

Accenture. 2016. Building digital trust: The role of data ethics in the digital age. Accenture Labs.

Glenn Adams. 2014. Decolonizing methods: African studies and qualitative research. *Journal of social and personal relationships*, 31(4):467–474.

William G Axinn. 1991. The influence of interviewer sex on responses to sensitive questions in nepal. *Social Science Research*, 20(3):303–318.

Rosaline Barbour. 2013. *Introducing qualitative research: a student's guide*. Sage.

Marie-Louise Barry, Herman Steyn, and Alan Brent. 2008. Determining the most important factors for sustainable energy technology selection in africa: Application of the focus group technique. In *PICMET'08-2008 Portland International Conference on Management of Engineering & Technology*, pages 181–187. IEEE.

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.

Nicole Bergen and Ronald Labonté. 2020. "everything is perfect, and we have no problems": Detecting and limiting social desirability bias in qualitative research. *Qualitative Health Research*, 30(5):783–792.

Richard A Berk. 1983. An introduction to sample selection bias in sociological data. *American sociological review*, pages 386–398.

The Nuffield Council on Bioethics. 2002. The ethics of research related to healthcare in developing countries. The Nuffield Council on Bioethics is funded jointly by the Medical Research Council, the Nuffield Foundation and the Wellcome Trust.

Christopher Blattman, Horacio Larreguy, Benjamin Marx, and Otis R Reid. 2019. Eat widely, vote wisely? lessons from a campaign against vote buying in uganda. Technical report, National Bureau of Economic Research.

Alan Bryman. 2012. Mixed methods research: combining quantitative and qualitative research. In Alan Bryman, editor, *Social Reserach Methods*, forth edition, chapter 27, pages 628–652. Oxford University Press, New York.

Badru Bukenya, Sam Hickey, and Sophie King. 2012. Understanding the role of context in shaping social accountability interventions: towards an evidence-based approach. *Manchester: Institute for Development Policy and Management, University of Manchester*.

de Cenival M. 2008. Ethics of research: the freedom to withdraw. *Bulletin de la Societe de pathologie exotique*, 101(2):98—-101.

Luc J. Christiaensen and Kalanidhi Subbarao. 2005. Towards an understanding of household vulnerability in rural kenya. *Journal of African Economies*, 14(4):520–558.

Tom Clark. 2008. We're over-researched here!' exploring accounts of research fatigue within qualitative research engagements. *Sociology*, 42(5):953–970.

Costanza Conforti, Stephanie Hirmer, Dai Morgan, Marco Basaldella, and Yau Ben Or. 2020. Natural language processing for achieving sustainable development: the case of neural labelling to enhance community profiling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8427–8444, Online. Association for Computational Linguistics.

Steven S Coughlin. 1990. Recall bias in epidemiologic studies. *Journal of clinical epidemiology*, 43(1):87–91.

D. A. Diallo, O. K. Doumbo, C. V. Plowe, T. E. Wellems, E. J. Emanuel, and S. A Hurst. 2005. Community permission for medical research in developing countries. *Infectious Diseases Society of America*, 41(2):255—-259.

Nora Dudwick, Kathleen Kuehnast, Veronica N. Jones, and Michael Woolcock. 2006. Analyzing social capital in context: A guide to using qualitative methods and data.

Terry Dunbar and Margaret Scrimgeour. 2006. Ethics in indigenous research–connecting with community. *Journal of Bioethical Inquiry*, 3(3):179–185.

J Dyer, L.C Stringer, A.J Dougill, J Leventon, M Nshimbi, F Chama, A Kafwifwi, J.I Muledi, J.-M.K Kaumbu, M Falcao, S Muhorro, F Munyemba, G.M Kalaba, and S Syampungani. 2014. Assessing participatory practices in community-based natural resource management: Experiences in community engagement from southern africa. *Journal of Environmental Management*, 137:137–145.

Ilker Etikan, Sulaiman Abubakar Musa, and Rukayya Sunusi Alkassim. 2016. Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics*, 5(1):1–4.

FIDH. 2012. Women's rights in Uganda: gaps between policy and practice. Technical report, International Federation for Human Rights, Paris.

Bruce B Frey. 2018. *The SAGE encyclopedia of educational research, measurement, and evaluation*. Sage Publications.

Batya Friedman, Lisa P. Nathan, and Daisy Yoo. 2017. Multi-lifespan information system design in support of transitional justice: Evolving situated design principles for the long (er) term. *Interacting with Computers*, 29(1):80–96.

Morris Gallagher, Tim Hares, John Spencer, Colin Bradshaw, and Ian Webb. 1993. The nominal group technique: a research tool for general practice? *Family practice*, 10(1):76–81.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Thomas Gilovich, Dale Griffin, and Daniel Kahneman. 2002. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press.

Brett A Gleitsmann, Margaret M Kroma, and Tammo Steenhuis. 2007. Analysis of a rural water supply project in three communities in mali: Participation and sustainability. In *Natural resources forum*, volume 31, pages 142–150. Wiley Online Library.

Alemu Gonsamo and Petra D'Odorico. 2014. Citizen science: best practices to remove observer bias in trend analysis. *International journal of biometeorology*, 58(10):2159–2163.

Matthew J Gordon. 2019. Language variation and change in rural communities. *Annual Review of Linguistics*, 5:435–453.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online. Association for Computational Linguistics.

Daniel Hammett and Deborah Sporton. 2012. Paying for interviews? negotiating ethics, power and expectation. *Area*, 44(4):496–502.

Anna Harding, Barbara Harper, Dave Stone, Catherine O'Neill, Patricia Berger, Stuart Harris, and Jamie Donatuto. 2012. Conducting research with tribal communities: Sovereignty, ethics, and data-sharing issues. *Environmental health perspectives*, 120(1):6–10.

J. Hentschel. 1999. Contextuality and data collection methods: A framework and application to health service utilisation. *Journal of Development Studies*, 35(4):64–94.

Ursula Hess, RB Jr Adams, and Robert E Kleck. 2009. Intergroup misunderstandings in emotion communication. *Intergroup misunderstandings: Impact of divergent social realities*, pages 85–100.

Stephanie Hirmer. 2018. *Improving the Sustainability of Rural Electrification Schemes: Capturing Value for Rural Communities in Uganda*. Ph.D. thesis, University of Cambridge.

Stephanie Hirmer and Peter Guthrie. 2016. Identifying the needs of communities in rural uganda: A method for determining the 'user-perceived value'of rural electrification initiatives. *Renewable and Sustainable Energy Reviews*, 66:476–486.

Stephanie Hirmer and Peter Guthrie. 2017. The benefits of energy appliances in the off-grid energy sector based on seven off-grid initiatives in rural uganda. *Renewable and Sustainable Energy Reviews*, 79:924–934.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

D. O. Irabor and P. Omonzejele. 2009. Local attitudes, moral obligation, customary obedience and other cultural practices: their influence on the process of gaining informed consent for surgery in a tertiary institution in a developing country. *Developing world bioethics*, 9(1):34—-42.

ITU. 2019. International Telecommunication Union World Telecommunication/ICT Indicators Database. World Bank Open Data.

Stephany Jimenez. 2020. Creatively Communicating through Visual and Verbal Art- Poetry and Murals. Yale National Initiative.

Fridah Katushemererwe and John Nerbonne. 2015. Computer-assisted language learning (call) in support of (re)-learning native languages: the case of runyakitara. *Computer Assisted Language Learning*, 28(2):112–129.

Benjamin Philip King. 2015. *Practical Natural Language Processing for Low-Resource Languages.* Ph.D. thesis, University of Michigan.

Dessi Kirilova and Sebastian Karcher. 2017. Rethinking data sharing and human participant protection in social science research: Applications from the qualitative realm. *Data Science Journal*, 16.

Jenny Kitzinger. 1994. The methodology of Focus Groups: the importance of interaction between research participants. *Sociology of Health and Illness*, 16(1):103–121.

Krishna Kumar. 1987. *Conducting group interviews in developing countries*. US Agency for International Development Washington, DC.

Rachid Laajaj and Karen Macours. 2017. *Measuring skills in developing countries*. The World Bank.

Alain Labrique, Emily Blynn, Saifuddin Ahmed, Dustin Gibson, George Pariyo, and Adnan A Hyder. 2017. Health surveys using mobile phones in developing countries: automated active strata monitoring and other statistical considerations for improving precision and reducing biases. *Journal of medical Internet research*, 19(5):e121.

Emily A Largent and Holly Fernandez Lynch. 2017. Paying research participants: regulatory uncertainty, conceptual confusion, and a path forward. *Yale journal of health policy, law, and ethics*, 17(1):61.

A. Launiala and T. Kulmala. 2006. The importance of understanding the local context: Women's perceptions and knowledge concerning malaria in pregnancy in rural malawi. *Acta Tropica*, 98(2):111–117.

Paul J Lavrakas. 2008. *Encyclopedia of survey research methods*. Sage Publications.

Tuan Anh Le, David Moeljadi, Yasuhide Miura, and Tomoko Ohkuma. 2016. Sentiment analysis for low resource languages: A study on informal indonesian tweets. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 123–131.

KM MacQueen, E McLellan, DS Metzger, S Kegeles, RP Strauss, and et al. 2001. What is community? an evidence-based definition for participatory public health. *American Journal of Public Health*, 91:1929–1938.

Eric M. Mafuta, Lisanne Hogema, Thérèse N.M. Mambu, Pontien B Kiyimbi, Berthys P. Indebe, Patrick K. Kayembe, Tjard De Cock Buning, and Marjolein A. Dieleman. 2016. Understanding the local context and its possible influences on shaping, implementing and running social accountability initiatives for maternal health services in rural democratic republic of the congo: a contextual factor analysis. *Science Advances*, 16(1):1–13.

B. Mallick, K. Rubayet Rahaman, and J. Vogt. 2011. Social vulnerability analysis for sustainable disaster mitigation planning in coastal bangladesh. *Disaster Prevention and Management*, 20(3):220–237.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. *arXiv preprint arXiv:2003.04986*.

M N Marshall. 1996. Sampling for qualitative research. *Family practice*, 13(6):522–5.

Salule J. Masangwi. 2015. Methodology for Solar PV Needs Assessment in Chikwawa, Southern Malawi. Technical report, Malawi Renewable Energy Acceleration Programme (MREAP) MREAP: Renewable Energy Capacity Building Programme (RECBP) Produced.

Joseph D Matarazzo, Morris Weitman, George Saslow, and Arthur N Wiens. 1963. Interviewer influence on durations of interviewee speech. *Journal of Verbal Learning and Verbal Behavior*, 1(6):451–458.

Keith McAdam. 2004. The ethics of research related to healthcare in developing countries. *Acta Bioethica*, 10(1):49–55.

J. R. McMillan, C. Conlon, and Nuffield Council on Bioethics. 2004. The ethics of research related to healthcare in developing countries. *Journal of medical ethics*, 30:204–206.

Albine Moser and Irene Korstjens. 2018. Series: Practical guidance to qualitative research. part 3: Sampling, data collection and analysis. *European Journal of General Practice*, 24(1):9–18.

Aaron Mueller, Garrett Nicolai, Arya D McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. An analysis of massively multilingual neural machine translation for low-resource languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3710–3718.

David Musoke, Charles Ssemugabo, Rawlance Ndejjo, Sassy Molyneux, and Elizabeth Ekirapa-Kiracho. 2020. Ethical practice in my work: community health workers' perspectives using photovoice in wakiso district, uganda. *BMC Medical Ethics*, 21(1):1–10.

Kevin L. Nadal. 2017. Sampling Bias and Gender. In *The SAGE Encyclopedia of Psychology and Gender*.

Judith Nakayiza. 2016. The sociolinguistic situation of english in uganda. *Ugandan English. Amsterdam/Philadelphia: John Benjamins*, pages 75–94.

N. Narayanasamy. 2009. *Participatory Rural Appraisal: Principles, Methods and Application*, first edition. SAGE Publications India Pvt Ltd, New Dehli.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

V. Nikulina, H. Larson Lindal, J.and Baumann, D. Simon, and H Ny. 2019. Lost in translation: A framework for analysing complexity of co-production settings in relation to epistemic communities, linguistic diversities and culture. *Futures*, 113(102442):1–13.

OECD. 2019. Survey of refugees and humanitarian staff in Uganda. Joint effort by Ground Truth Solutions (GTS) and the Organisation for Economic Co-operation and Development (OECD) Secretariat with financial support from the United Kingdom's Department for International Development (DFID).

Daniel G Oliver, Julianne M Serovich, and Tina L Mason. 2005. Constraints and opportunities with interview transcription: Towards reflection in qualitative research. *Social forces*, 84(2):1273–1289.

Sean Packham. 2016. *Crowdsourcing a text corpus for a low resource language*. Ph.D. thesis, University of Cape Town.

Christopher J Pannucci and Edwin G Wilkins. 2010. Identifying and avoiding bias in research. *Plastic and reconstructive surgery*, 126(2):619.

Erin S. Parcell and Katherine A. Rafferty. 2017. Interviews, recording and transcribing. In Mike Allen, editor, *The SAGE Encyclopedia of Communication Research Methods*, pages 800–803. SAGE Publications, Thousand Oaks.

Sonny S Patel, Rebecca K Webster, Neil Greenberg, Dale Weston, and Samantha K Brooks. 2020. Research fatigue in covid-19 pandemic and post-disaster research: causes, consequences and recommendations. *Disaster Prevention and Management: An International Journal*.

Uwe Peters. 2020. What is the function of confirmation bias? *Erkenntnis*, pages 1–26.

Stephanie Carroll Rainie, Tahu Kukutai, Maggie Walter, Oscar Luis Figueroa-Rodríguez, Jennifer Walker, and Per Axelsson. 2019. Indigenous data sovereignty.

Thiagarajan Ramakrishnan, Mary C. Jones, and Anna Sidorova. 2012. Factors influencing business intelligence (bi) data collection strategies: An empirical investigation. *Decision Support Systems*, 52:486—-496.

Max Roser and Esteban Ortiz-Ospina. 2016. Literacy. *Our World in Data*.

Dwaipayan Roy, Sumit Bhatia, and Prateek Jain. 2020. A topic-aligned multilingual corpus of Wikipedia articles for studying information asymmetry in low resource languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2373–2380, Marseille, France. European Language Resources Association.

Mary Kathryn Salazar. 1990. Interviewer bias: How it affects survey research. *Aaohn Journal*, 38(12):567–572.

Inaad Mutlib Sayer. 2013. Misunderstanding and language comprehension. *Procedia-Social and Behavioral Sciences*, 70:738–748.

Deven Shah, H. Andrew Schwartz, and Dirk Hovy. 2019. Predictive biases in natural language processing models: A conceptual framework and overview.

David Silverman. 2013. *Doing Qualitative Reserach*, fourth edition. SAGE Publications Inc.

Jonathan Skinner et al. 2013. *The interview: An ethnographic approach*, volume 49. A&C Black.

David Stewart, Prem Shamdasani, and Dennis Rook. 2007. Group dynamics and focus group research. In David Stewart, Prem Shamdasani, and Dennis Rook, editors, *Focus Groups: Theory and Practice*, chapter 2, pages 19–36. SAGE Publications, Thousand Oaks.

Deryn Strange and Melanie KT Takarangi. 2015. Memory distortion for traumatic events: the role of mental imagery. *Frontiers in psychiatry*, 6:27.

Latanya Sweeney. 2000. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34.

Fasil Tekola, Susan J. Bull, Farsides Bobbie, J. Newport Melanie, Adeyemo Adebowale, N. Rotimi Charles, and Davey Gail. 2009. Tailoring consent

to context: Designing an appropriate consent process for a biomedical study in a low income setting. *PLOS Neglected Tropical Diseases*, 3:7.

Lisa Thorley and Emma Henrion. 2019. DFID ethical guidance for research, evaluation and monitoring activities. Prepared for the UK Department for International Development.

Paulina O. Tindana, Linda Rozmovits, Renaud F. Boulanger, Sunita V.S. Bandewar, Raymond A. Aborigo, Abraham V.O. Hodgson, Pamela Kolopack, and James V. Lavery. 2011. Aligning community engagement with traditional authority structures in global health research: A case study from northern ghana. *American Journal of Public Health*, 101:1857–1867.

Paulina O. Tindana, Jerome A. Singh, C. Shawn Tracy, Ross E.G. Upshur, Abdallah S. Daar, Peter A. Singer, Janet Frohlich, and James V. Lavery. 2007. Grand challenges in global health: Community engagement in research in developing countries. *PLOS Medicine*, 4(9).

Marc-Adèlard Trembley. 1957. The Key Informant Technique: A Nonethnographic Application. Technical report, Cornell University, New York.

Angela K. Troyer. 2011. *Serial Position Effect*, pages 2263–2264. Springer New York, New York, NY.

J. W. Tukey. 1980. We need both exploratory and confirmatory. *American Statistician*, 34(1):23–25.

UKRI. 2020a. Data protection Guidance. Economic and Social Research Council.

UKRI. 2020b. Do participants have a right to withdraw consent? Economic and Social Research Council.

UNDESA. 2018. United Nations Department of Economic and Social Affairs, Population Division: World Urbanization Prospects. World Bank Open Data.

UNESCO. 2018. UNESCO Institute for Statistics Adult Literacy Rate. World Bank Open Data.

UNHCR. 2020. Refugees and Nationals by District. Uganda Comprehensive Refugee Response Portal.

Melissa Upjohn and Kimberly Wells. 2016. Challenges associated with informed consent in low-and low-middle-income countries. *Frontiers in veterinary science*, 3:92.

USAID. 2006. Powering Health, Electrification Options for Rural Health Centers. Technical report, USAID, Washington DC, USA.

Ana Maria de Albuquerque Vasconcellos and Mário Vasconcellos Sobrinho. 2014. Knowledge and culture: two significant issues for local level development programme analysis. *Interações (Campo Grande)*, 15(2):285–300.

Martijn Veltkamp, Ruud Custers, and Henk Aarts. 2011. Motivating consumer behavior by subliminal conditioning in the absence of basic needs: Striking even while the iron is cold. *Journal of Consumer Psychology*, 21(1):49–56.

Inga Vesper. 2019. Facts & Figures: Disabilities in developing countries. Sci Dev Net.

Robert S Walker and Marcus J Hamilton. 2011. Social complexity and linguistic diversity in the austronesian and bantu population expansions. *Proceedings of the Royal Society B: Biological Sciences*, 278(1710):1399–1404.

Ke Wang, Steven Goldstein, Madeleine Bleasdale, Bernard Clist, Koen Bostoen, Paul Bakwa-Lufu, Laura T Buck, Alison Crowther, Alioune Dème, Roderick J McIntosh, et al. Ancient genomes reveal complex patterns of population movement, interaction, and replacement in sub-saharan africa.

Christopher D Watkins, Lisa M DeBruine, Anthony C Little, David R Feinberg, and Benedict C Jones. 2012. Priming concerns about pathogen threat versus resource scarcity: dissociable effects on women's perceptions of men's attractiveness and dominance. *Behavioral Ecology and Sociobiology*, 66(12):1549–1556.

Roger P Winter. 1983. Uganda-creating a refugee crisis. *Cultural Survival Quarterly*, 7(2).

Philip Woodhouse. 1998. Thinking with People and Organizations. In Alan Thomas, Joanna Chataway, and Marc Wuyts, editors, *Finding out Fast: Investigative Skills for Policy and Development*, first edition, chapter Part III, pages 127–146. SAGE Publications Inc, Milton Keynes.

B. Wu. 2014. Embedding research in local context: local knowledge, stakeholders' participation and fieldwork design. *Field Research Method Lab at LSE*.

Derek Yach. 1992. The use and value of qualitative methods in health research in developing countries. *Social science & medicine*, 35(4):603–612.