Data set name: Demographic-Rich Qualitative UPV-Interviews Dataset (DR-QI)

Data set developers: Josephine Tumwesige, Yau Ben-Or, Yair Perry, Vishal Narula, Stephanie Hirmer, Costanza Conforti, Goya van Boven

Data statement authors: Stephanie Hirmer , Costanza Conforti, Goya van Boven

Contacts: info@ruralsenses.com

A. Curation Rationale

The DR-QI dataset contains 5,333 sentences. The sentences constitute excerpts from 214 interviews, conducted in 56 rural villages in India and 68 villages in Uganda. The exact names of the villages in which interviews were collected are not released in order to protect the identity of participants. The interviews were part of a larger project to develop an impact framework for off-grid energy appliances in Low- and Middle-Income countries (LMICs).

The sentences in the DR-QI dataset were annotated with the User-Perceived Value (UPV) approach. The dataset contains additional demographic information for each speaker. The dataset releases the 5,333 translated sentences, along with their UPV annotations. For each speaker, ten self-reported categorical demographic features are included.

B. Language Variety and Translator Demographic

The sentences in the DR-QI dataset are in English. Sentences were translated from original transcriptions in 7 local languages.

To ensure translators could take area and tribe specific linguistic nuances into account, all translators were from the regions in which interviews were carried out.

C. Speaker Demographic

Details on the speakers' demographics are reported in Table 1. For each speaker the following self-reported demographic features are included: age, gender, marital status, disability (yes/no), education level, occupation, household size, number of children, income relative to the participants group from the same country, and absolute poverty status. All demographic features are categorical.

Total speakers	214
Gender	Equal split women/men
Nationality	Mostly Ugandan and Indian
Age	Between 18 and 77.
Socio Economic Status	Variable: political and religious leaders excluded, as well as close family members.

Table 1: Speakers Demographic in the DR-QI dataset.

D. Annotator Demographic

The 8 annotators all lived in LMICs and were familiar with the local context of the interviews. Prior to starting the annotation process, annotators received a training workshop and took part in a short quiz. Annotations were carried out on a user-friendly platform. Annotators received a competitive salary that matched the local context.

E. Speech Situation

All interviews were collected between 2020 and 2021. The interviews were conducted in locations familiar to the interviewees, mostly open air. The interviews were conducted both individually and in groups of 6 people following standard focus group methods. To avoid direct inquiry, the interviews were conducted by means of the *UPV game*, which is described in detail in Hirmer and Guthrie (2016), resulting in semi-structured interactions. Even if the speech situation can be characterized as a dialogue, the interviewers were instructed to talk as little as possible, in order to avoid external influence, and to elicit answers to simple interactions such as *why probing*.

As a consequence, many of the interviews may be better characterized as monologues.

F. Text Characteristics

All translators were bilingual, but English wasn't the native language of any of them. As a consequence, the DR-QI dataset contains some grammatical errors. Moreover, the DR-QI dataset contains many examples of oral constructions. All proper nouns (people, tribes, locations, ...) in the DR-QI dataset are anonymized with a special tag.

G. Recording Quality

Original recordings aren't released.