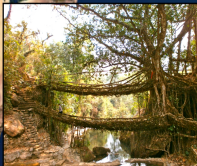


# Sharing Linguistic Data:

Social Science concerns and audio data anonymisation

Hiram Ring

[www.hiramring.com](http://www.hiramring.com)





# What is Linguistic data?

- speech-related data recorded from humans
- transcribed text, or text from literary sources
- annotations of any of the above
- transformations of any of the above (ETC...)



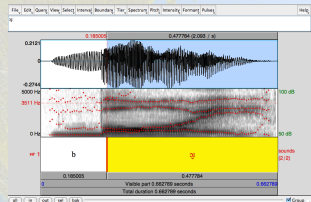
**kaba**

maize, corn, cobs, cawcaw

maize  
cobs  
cawcaw

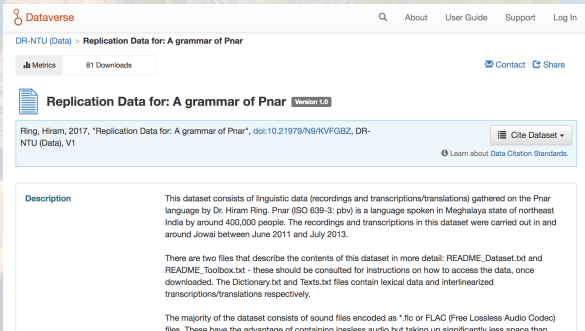
Translate



Word	Morphemes	Lex. Entries	Lex. Gloss	Lex. Gram. Info	Word Cat.
yalo	pus	pus	green	ns	nom
yalo	pus	pus	green	ns	nom
yalo	pus	pus	green	ns	nom

# What are the concerns?

- permission to record/share data
- acknowledgement/attribution of sources
- anonymization of sources if necessary
- license/availability/use of data
- quality of released data (processing, replication)



The screenshot shows the Dataverse interface for a dataset titled "Replication Data for: A grammar of Pnar". The page includes a search bar, navigation links (About, User Guide, Support, Log In), and a breadcrumb trail: "DR-NTU (Data) > Replication Data for: A grammar of Pnar". Below the title, there are buttons for "Metrics" (showing 81 Downloads) and "Share", along with "Contact" and "Share" icons. A document icon is next to the title, which also includes a "Version 1.0" label. The citation information is displayed as: "Ring, Hiram, 2017, 'Replication Data for: A grammar of Pnar', doi:10.21979/N9/KVFBGZ, DR-NTU (Data), V1". A "Cite Dataset" button and a link to "Learn about Data Citation Standards" are also present. The "Description" section contains three paragraphs of text.

**Description**

This dataset consists of linguistic data (recordings and transcriptions/translations) gathered on the Pnar language by Dr. Hiram Ring. Pnar (ISO 639-3: pbv) is a language spoken in Meghalaya state of northeast India by around 400,000 people. The recordings and transcriptions in this dataset were carried out in and around Jowai between June 2011 and July 2013.

There are two files that describe the contents of this dataset in more detail: README\_Dataset.txt and README\_Toolbox.txt - these should be consulted for instructions on how to access the data, once downloaded. The Dictionary.txt and Texts.txt files contain lexical data and interlinearized transcriptions/translations respectively.

The majority of the dataset consists of sound files encoded as \*.flc or FLAC (Free Lossless Audio Codec) files. These have the advantage of containing lossless audio but taking up significantly less space than

# Anonymization of audio data

– involves editing the shared sound file (a copy of the original) using an editor such as Audacity

