DECODIS

Our ASR study for Swahili and Yoruba released on GitHub

May 2025

For the past five years, Decodis has been collecting voice data in low-resource languages, at scale.

Decodis gathers open-ended spoken responses over automated phone call (IVR). Examples of projects funded by the **Gates Foundation**:

- Digital Portfolios of the Poor
- Digital Personas

See the advantages of collecting voice data with IVR as well as our consent process.

People answer questions in natural language

They use pauses, ums and ahs, shortcuts, local expressions, and switch between local language and English.

Click the play button to hear an example.

Natural language training significantly improves performance against real-world data, even with smaller corpora

Our **Swahili** ASR results based on natural language

Model used

We used the Whisper Large v2 architecture with identical settings for fair comparison. The base model is reported with no training as a benchmark.

Three sets of training data

We **trained** this model on:

- Open-source data[1]: 400 hours
- Decodis natural language dataset: 50 hours
- Combined: 450 hours

Two test sets

We tested against two test data sets:

- Clean benchmark: FLEURS-created with actors reading text.
- Real world: Decodis-created with IVR with natural speech, showcasing flaws.

	Swahili Word Error Rates (WER)[2]		
	Clean benchmark	**Real world**	**Results**
Whisper Large v2 (no training)	39%	100%	Model does not have enough representation for Swahili in real settings.
Model trained on open-source data	14%	70%	Model collapses on real world data and does not recognize semantic patterns for generalization. This showcases open-source datasets do not capture real world nuances.
Model trained on Decodis natural language	35%	46%	The model trained with our data learns the language as it generalizes across different accents as well as flaws in different settings.
Combined	13%	43%	Combined model shows ability to learn language, even with a small proportion of natural language.

[1] Open-source data, OpenSLR (Project ALFFA).

[2] The ratio of number of misclassified words to the total number of words in the audio. Lower is better.

Very poor performance but natural language training again improves performance against real world data

Our **Yoruba** ASR results based on natural language

Model used

We used the Whisper Large v2 architecture with identical settings for fair comparison. The base model is reported with no training as a benchmark.

Three sets of training data

We **trained** this model on:

- Open-source data[1]: 45 hours
- Decodis natural language dataset: 25 hours
- Combined: 70 hours

Two test sets

We tested against two test data sets:

- Clean benchmark: FLEURS created with actors reading text.
- Real world: Decodis created with IVR in a range of settings and domains, showcasing flaws.

	Yoruba Word Error Rates (WER)[2]		
	Clean benchmark	**Real world**	**Results**
Whisper Large v2 (no training)	100%	100%	Model had no representation as the base dataset did not consist of Yoruba.
Model trained on open-source data	26%	81%	Open-source model cannot generalize and overfits to clean speech.
Model trained on Decodis natural language	66%	65%	The model trained with our natural data learns the language as it generalizes across different accents as well as flaws in different settings, providing similar performances in both.
Combined	25%	62%	Combined model shows ability to learn language, even with small data set.

[1] Open-source data, OpenSLR (SLR68).

[2] The ratio of number of misclassified words to the total number of words in the audio. Lower is better.

Hard-to-access natural language data at scale, and at a low cost.

Powerful tool for training models

Scale

Automation allows us to interview **1,000+ people in one day.**

Low cost

Each interview costs the same as a phone call.

Broad range of dialects and accents

Our surveys reach a broad range of populations for each language, including those who are illiterate and can only understand and speak their own language.

Click here to read more about how we obtain consent.

Fosters the best environment for natural language

Easy to access

No need to download an app. They only need a **basic phone** and to use their voice and their keypad.

Welcoming

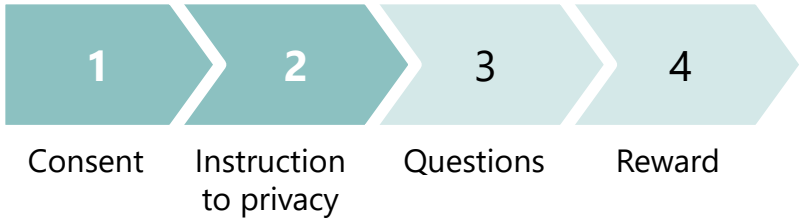
In-country voice actors with the right dialects and accents ask questions in local language in a friendly way. **Long responses** are encouraged.

Equitable

Respondents feel like in a conversation, not an interview. They speak longer and more meaningfully.[1]

[1]Collins et al, Sentiment of Bangladeshi Residents Toward Covid-19 Lockdowns: Qualitative Analyses of Open-Ended Responses in a Large Panel Survey. BRAC University, 2023. Available here.

Obtaining consent



Each IVR call first plays this consent script in local language.

Participants respond via keypad entry either giving consent or choosing not to participate.

Consent language

"I am part of an international survey company called Decodis. I am contacting you to ask if you would answer some questions in our language. The reason why I want your help is so we can help people who speak different languages speak to each other. We appreciate your answering some questions about health and agriculture so we can record your voice talking about these subjects. But your responses will remain anonymous, and no one will be able to tie your responses back to your identity. The interview should take up no more than 20 minutes, and at the end of the interview, we will send you [amount] *as a token of our appreciation for your time. It's your decision, and there are no consequences to saying no. If you want to stop the survey or skip a question, that's ok too.*

Are you comfortable doing this interview? Press 1 for Yes, Press 2 for No, and Press 3 if you would like me to call you back at a different time."

After pressing 1

Thank you! Can you go into a quiet place to do the survey so I can hear your answers? Please just talk into the phone as long as you want.

DECODIS

Social Research. Reimagined.

info@decodis.com

www.decodis.com