Griots Speech Dataset	The Griots Dataset is a combined Bambara audio, and semi-professionally transcribed, and translated into French speech dataset. This dataset was developed with the intention to build a Bambara ASR model.
DATASET LINK	DATA CARD AUTHOR(S)
https://github.com/robotsmali-ai/jeli-asr https://zenodo.org/record/7296317	Sebastien Diarra, RobotsMali: Owner

Authorship

Publishers

PUBLISHING ORGANIZATION	INDUSTRY TYPE(S)	CONTACT DETAIL(S)
RobotsMali	Not-for-profit - Academic - Tech	Publishing POC: Michael Leventhal Affiliation: RobotsMali Contact: mleventhal@robotsmali.org Website: https://robotsmali.org

Dataset Owners

TEAM(S)	CONTACT DETAILS	AUTHOR(S)
	Dataset Owner(s): RobotsMali Affiliation: RobotsMali	Michael Leventhal, PI

<table>
<tr><td></td><td>Contact:
research@robotsmali.org
Website:
https://ai.robotsmali.org</td><td>Sebastien Diarra, Manager
Mouktar Traore, Supervisor
Alou Dembele, Supervisor</td></tr>
<tr><td>Funding Sources</td><td></td><td></td></tr>
<tr><td>INSTITUTION(S)</td><td colspan="2">FUNDING OR GRANT SUMMARY(IES)</td></tr>
<tr><td>GOOGLE, LLC</td><td colspan="2">Grant provided to develop a functioning ASR model in Bambara; Also to improve the performance of an existing Bambara-French Translation Model.</td></tr>
<tr><td>Dataset Overview</td><td></td><td></td></tr>
<tr><td>DATA SUBJECT(S)</td><td>DATASET SNAPSHOT</td><td>CONTENT DESCRIPTION</td></tr>
<tr><td>Cultural Data
Social Commentary
History
Narrations</td><td><table>
<tr><td>Size</td><td>16 GB</td></tr>
<tr><td>Length</td><td>30 hours</td></tr>
<tr><td>Utterances (Clips)</td><td>29800</td></tr>
<tr><td>Ave. Clips Length</td><td>3.02 s</td></tr>
<tr><td>Tokens</td><td>300923</td></tr>
<tr><td>Types</td><td>62753</td></tr>
<tr><td>M/F Speaker Ratio</td><td>23/7</td></tr>
</table></td><td>The dataset consists of 1 hour recording of 30 griots. Each recording is focused on the historical and cultural records of Mali.</td></tr>
<tr><td colspan="3">Dataset Version and Maintenance</td></tr>
<tr><td>MAINTENANCE STATUS</td><td>VERSION DETAILS</td><td>MAINTENANCE PLAN</td></tr>
<tr><td>Actively Maintained</td><td>Current Version: 1.0.0
Last Updated: 09/2022
Release Date: 08/2022</td><td>The repository is adjusted accordingly, to augment the data quality. Additional data in the domain is sought after.

Versioning: Major.Minor.Patch.</td></tr>
</table>

		Semantic Versioning specification is adopted. Updates: This will occur on a cyclical basis, as we expect to increase the dataset with related works. Errors: Errors are addressed right when they are flagged. https://github.com/robotsmali-ai/jeli-asr/issues Feedback: Contact the manager, send us an email at research@robotsmali.org.
Example of Data Points		
PRIMARY DATA MODALITY	SAMPLING OF DATA POINTS	DATA FIELDS
Text Data Audio Data	https://github.com/RobotsMali-AI/jeli-asr/blob/master/README.md	Utterance start (milliseconds) Utterance end (milliseconds) Utterance transcription Utterance translation Second elapsed Recording Fragment ID
TYPICAL DATA POINT		
[4640, 6460, "Alhamdulilayi Rabil Alamiina", "Alhamduolilaye Rabil Alamina!", 1.82, "griots_r1_1"]		
Motivations & Intentions		
PURPOSE(S)	DOMAIN(S) OF APPLICATION	Intended And/Or Suitable Use Case(s)
Safe for research use Safe for production use Training Inferencing	Machine Learning, Automatic Speech Recognition, Machine Translation	- Bringing the bambaraphone communities online, through cultural data.

		- Resource generation for a low-resource language.
Provenance		
Collection		
METHOD (S) USED	METHODOLOGY DETAIL(S)	SOURCE DESCRIPTION (S)
Crowdsourced - Paid	Audio Recording Primary modality of collected data: Audio Data Transcription/Translation Platform: ELAN Primary modality of collected data: Text Data Dates of collection: 05 2022 - 08 2022	Audio: Griots. Oral stories recorded with professional recording equipment. Text: Linguists / Transcribers, using ELAN
Use in ML or AI Systems		
Dataset Use (s)	Notable Features	Usage Example
Training Testing Validation Fine Tuning	Tokens Types Utterances	https://github.com/RobotsMali-AI/jeli-asr/tree/master/asr
Annotations & Labeling		
ANNOTATION WORKFORCE TYPE		
Human Annotations (Non-Expert) Human Annotations (Contractors) Human Annotations (Crowdsourcing)		

Human Annotators		
	ANNOTATOR DESCRIPTION	ANNOTATOR TASK
	Transcription / Translation Use ELAN to transcribe audio fragments. Number: 20 persons team Level: Amateur linguist Editors Edited transcriptions and translations. Number: 5 persons team Level: Experts Language Distribution of annotators: - French: 3 - Bambara: 2	Transcription / Translation: Transcribing audio fragments using ELAN and translating. Editors Take the ELAN file from the transcription step
LICENSING		
LICENSE TYPE(S)	LICENSE BREAKDOWN	LICENSE PERMISSIONS
CC-BY-SA 4.0	This work is licensed under the Creative Commons Attribution 4.0 International License. CC-BY-SA 4.0	• Share — copy and redistribute the material in any medium or format • Adapt — remix, transform, and build upon the material for any purpose, even commercially. • Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any

		reasonable manner, but not in any way that suggests the licensor endorses you or your use.

RobotsMali

Centre National Collaboratif pour l'Éducation en Robotique et en Intelligence Artificielle

Ecole Normale d'Enseignement Technique et Professionnel (ENETP)

Cité-Universitaire de Kabala

Bamako, Mali